THE UNIVERSITY
of EDINBURGH

Visual Context for Verb Sense Disambiguation and Multilingual Representation Learning

Spandana Gella

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2019

Abstract

Every day billions of images are uploaded to the web. To process images at such a large scale it is important to build automatic image understanding systems. An important step towards understanding the content of the images is to be able to understand all the objects, scenes and actions depicted in the image. These systems should be capable of integrating with natural language or text to be able to query and interact with humans for tasks such as image retrieval.

Verbs play a key role in the understanding of sentences and scenes. Verbs express the semantics of an actions as well as the interactions between objects participating in an event. Thus understanding verbs is central to both language and image understanding. However, verbs are known for their variability in meaning with context. Many studies in psychology have shown that contextual information plays an important role in semantic understanding and processing in the human visual system. We use this as intuition and understand the role of textual or visual context in tasks that combine language and vision.

Our research presented in this thesis focuses on the problems of integrating visual and textual contexts for: (i) automatically identifying verbs that denote actions depicted in the images; (ii) fine-grained analysis of how visual context can help disambiguate different meanings of verbs in a language or across languages; (iii) the role played by the visual and multilingual context in learning representations that allow us to query information across modalities and languages.

First, we propose the task of visual sense disambiguation, an alternative way of addressing the action recognition task. Instead of identifying the actions directly, we develop a two step process: identifying the verb that denotes the action being depicted in an image and then disambiguate the meaning of the verb based on the visual and textual context associated with the image. We first build a image-verb classifier based on the weak signal from image description data and analyse the specific regions that model focuses on while predicting the verb. We then disambiguate the meaning of the verb shown in the image using image features and sense-inventories. We test the hypothesis that visual and textual context associated with the image contribute to the disambiguation task.

Second, we ask whether the predictions made by such models correspond to human intuitions about visual verbs or actions. We analyse whether the image regions a verb prediction model identifies as salient for a given verb correlate with the regions fixated

by human observers performing an action classification task. We also compare the correlation of human fixations against visual saliency and center bias models.

Third, we propose the crosslingual verb disambiguation task: identifying the correct translation of the verb in a target language based on visual context. This task has the potential to resolve lexical ambiguity in machine translation when the visual context is available. We propose a series of models and show that multimodal models that fuse textual information with visual features have an edge over text or visual only models. We then demonstrate how visual sense disambiguation can be combined with lexical constraint decoding to improve the performance of a standard unimodal machine translation system on image descriptions.

Finally, we move on to learn joint representations for images and text in multiple languages. We test the hypothesis that context provided as visual information or text in other language contributes to better representation learning. We propose models to map text from multiple languages and images into a common space and evaluating the usefulness of the second language in multimodal search and usefulness of image in the crosslingual search. Our experiments suggest that exploiting multilingual and multimodal resources can help in learning better semantic representations that are useful for various multimodal natural language understanding tasks.

Our experiments on visual sense disambiguation, sense disambiguation across languages, multimodal and cross-lingual search demonstrate that visual context alone or combined with textual context is useful for enhancing multimodal and crosslingual applications.

Lay Summary

An ultimate goal of artificial intelligence is to create a system that can learn and understand information from vast amounts of data, communicate with the user and perform an action according to its user's instructions both from language and visual signal. A very crucial part of this is to be able to recognise and understand information. Verbs play a critical role in the semantic understanding of both language and visual information. The semantics of a verb in a scene changes with objects participating in an activity and the interactions between the objects. An effective way of representing and understanding verbs is necessary for systems to interpret the meaning of instructions to perform equal to humans. For example *playing guitar* vs. *playing football*. In this thesis, we formulate ways to identify verbs that denote actions depicted in images and demonstrate how visual context can help disambiguate different meanings of verbs. We examine whether visual context can help understanding mappings of meanings of verbs across languages. Finally, we study whether the visual and textual context is useful in learning representations that allow us to query information across modalities and languages.

Acknowledgements

I am deeply grateful to my supervisors, Frank Keller and Mirella Lapata, for their guidance, continuous support, encouragement and advice throughout my studies. I would like to thank Frank for giving me freedom to choose topics, explore directions and guiding me on finding right problems. Frank's feedback has improved the content and presentation of this thesis. Both Frank and Mirella played a great role in helping me choose research directions, guiding me on approaching and presenting research problems and for giving all the detailed feedback on the drafts of my papers. It was a great pleasure working with Mirella, I would like to thank for for all the enthusiasm and for her support.

I would like to thank my thesis committee members Lucia Specia and Timothy Hospedales for taking time to review my thesis, giving valuable feedback and improving various sections of the thesis. The work presented in this thesis is very relevant to the work of Lucia and I have learned a lot from reading her papers. I would like to thank Timothy Hospedales for being on my yearly committees and providing feedback on various chapters/papers presented in this thesis. It was great fun and learning experience collaborating with Desmond Elliot and Rico Sennrich. They were absolutely amazing. I would like to thank them for answering all my questions and teaching me all the great things. I learned a lot from both of you.

I am grateful to my masters supervisors Tim Baldwin and Paul Cook for encouraging me to apply for a PhD program. I was very fortunate to work with brilliant collaborators for both my internships during my PhD. Margaret Mitchell (Meg) is such an amazing advisor and a human being. She gave me immense support during and after my PhD. I would like to thank her for always being there for me and giving me all confidence and support during tough times. I would like to thank Mike Lewis and Marcus Rohrbach for giving me an amazing opportunity to intern at Facebook AI Research. I have met many wonderful researchers and colleagues.

I consider myself lucky to be part of ILCC. The Edinburgh NLP group meetings, reading groups have taught me how to read papers, appreciate research ideas, reviews ideas critically and most importantly how to present my work. I would like to thank all the group members who provided feedback on my PhD projects, presentations and research papers. I would like to thank all the faculty in Edinburgh NLP group for providing the fun and interactive environment.

I want to thank my colleagues and fellow PhD students for making my time in Edinburgh enjoyable . Special thanks to Annie, Carina, Chen, Dominikus, Des, Herman,

Lea, Michael, Rico, Helena for all the fun times, long lunch/dinner conversations. I would like to thank Clara, Irene, Lucia, Mona, Sameer, Gozde for being the best office mates, I could not have ask her better :). Special thanks to Siva, Bharat, Gangi, Praveen, Srikanth and Rupa for making Edinburgh another home, all the fun weekend getaways, food experiments, long board game/movie nights. I would like to thank all my friends in the bay area for helping me maintain a certain degree of sanity throughout the whole thesis writing phase.

I would like to thank Chris Manning for allowing me to work from Stanford, attend NLP group seminars and events when I visited Siva and while writing up my thesis. It has been an amazing experience to interact with StanfordNLP group. Special thanks to Vinod Prabhakaran for loaning his desk at Stanford.

I am grateful to my parents, in-laws and extended family for being a constant source of support and encouragement throughout my studies. I would like to thank my long term friend Lydia for always being there for me. I must acknowledge Siva without whose love, encouragement and continuous support, I would not have started or finished this thesis. I would like to dedicate this thesis to all the kind people who funded various stages of my studies, Siva and my family.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Spandana Gella*)

To Siva and my parents.

Table of Contents

Chapter 1

Introduction

An ultimate goal of artificial intelligence is to create a system that can process, learn and understand information from vast amounts of data to communicate with a user and perform an action according to the user's instructions. Recent advances in the technology, easy accessibility of smart devices equipped with a camera and the popularity of digital media and social networking have enabled people to upload and share images and videos which resulted in a steep rise of the visual content being uploaded to web daily. However, as the amount of data increases, the need for intelligent systems that can understand both language and visual information increases.

Recent studies in the computer vision community have successfully constructed models to efficiently recognize objects in the image or describe an image with language (Sermanet et al., 2013; Simonyan & Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Girshick et al., 2014; Karpathy & Li, 2015b). Despite significant advances in recognizing objects, one of the fundamental problems of computer vision is recognizing actions in images, which still remains challenging (Ronchi & Perona, 2015; Ramanathan et al., 2015; Pustejovsky et al., 2016). The semantics of verbs changes with respect to context: participants (objects) in an activity and interactions (relationships) between the participants. For example in Table 1.1 the verb *play* could be used to describe the action happening in both the images. In the first image, the verb *evokes* the meaning of playing a music instrument whereas in the second image it evokes the meaning of playing a sport. Most of the action recognition work so far can be viewed as a classification problem viz., labeling an image with a verb phrase label.

English	playing	playing
Spanish	tocando	jugando

Table 1.1: An example how visual information helps distinguish between two different usage of verb **play**

Most of the human object interaction datasets and models are targeted at images in a specific domain or limited to a set of labels and are far from applicable to real-world datasets and applications. This is not scalable as the number of target verbs and objects increase; the possible number of labels is the cartesian product of verb and object labels (Le et al., 2013b; Ronchi & Perona, 2015).

Our research presented in this thesis focuses on the problem of using language and linguistic resources to help understand images. We particularly study the problem of action recognition task by automatically identifying verbs that denote actions depicted in the images. We present a fine-grained analysis of how visual context can help disambiguate different meanings of verbs in a language or across languages. Additionally, we show the role played by visual and multilingual context in learning representations that allow us to query information across modalities and languages.

1.1 Contributions

This thesis makes following main contributions to the problem of automated understanding of verbs and actions in images and multimodal representation learning.

Visual Verb sense Disambiguation Task and Dataset: We introduce a new task of visual sense disambiguation for verbs and VerSe a new sense disambiguation datasets VerSe (short for Verb Sense) dataset consists a collection of images for 90 verbs annotated with their sense labels. VerSe images are also accompanied by other ground truth

annotations such as object annotations, action labels and descriptions.

We propose an unsupervised algorithm based on Lesk (Lesk, 1986) which performs visual sense disambiguation using textual, visual and multimodal information.

Cross-lingual Visual Sense Disambiguation Task and Dataset: We introduce the task of cross-lingual visual sense disambiguation (CLWSD) for verbs and MultiSense, a new crosslingual sense disambiguation dataset. MultiSense consists a collection of images for 55 verbs annotated with their translations in German and Spanish.

We propose a series of cross-lingual visual sense disambiguation models and show that multimodal models that fuse textual information with visual features perform best on the task. Additionally, we demonstrate that visual sense disambiguation can be used to improve the performance of a standard unimodal machine translation system on image descriptions.

Multilingual Multimodal Representation Learning We introduce novel models for learning representations of multiple languages and image in a joint space. Our model is formulated as a neural network architecture that learns bilingual multimodal space based on multiview learning of images and text in English and German. First, we show that the model yields representations useful for visual description task when presented with images only and image retrieval task when presented with text only.

Secondly, that our models are able to extend this behavior to languages other than English and can exploit information from other languages for better representation learning. Thirdly, that it can account for human behavior on semantic textual similarity tasks to rank how similar two given sentences are. Finally, we show that our models yield representations useful for cross-lingual search without using any parallel information between the languages.

1.2 Published Work

The contributions presented in this thesis are published in the following papers:

Chapter 2 was presented as:

> Gella, Spandana, Lapata, Mirella, and Keller, Frank. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In Proceedings

of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT.

Gella, Spandana and Keller, Frank. 2017. An analysis of action recognition datasets for language and vision tasks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

Gella, Spandana, Keller, Frank, and Lapata, Mirella. 2018. Disambiguating visual verbs. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Chapter 3 was presented as :

Gella, Spandana and Keller, Frank. An evaluation of image-based verb prediction models against human eye-tracking data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT.

Chapter 4 presented as:

Gella, Spandana, Elliot, Desmond and Keller, Frank. Cross-lingual Visual Sense Disambiguation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT.

Chapter 5 was presented as :

Gella, Spandana, Sennrich, Rico, Keller, Frank, and Lapata, Mirella. 2017. Image pivoting for learning multilingual multimodal representations. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

1.3 Thesis Outline

We present our visual sense disambiguation task and models proposed to address this task in Chapter 2. Chapter 3 describes our analysis on the correlation of human eye-tracking fixations vs. localizations of convolutional neural network models for action classification task. Chapter 4 introduces our model cross-lingual word sense disambiguation using visual context and its application to enhance Machine Translation models. In Chapter 5 we present models to learn joint space for images and text in multiple

languages using visual information as a bridge between languages, and conclude the thesis in Chapter 6.

Chapter 2 presents our annotation procedure to create our VerSe visual verb sense disambiguation dataset. We discuss how target verbs, senses and images were obtained. We explain how we use image-descriptions to generate weakly labeled data for training multi-label verb-classification models. We devise a multiple instance classification model that not only predicts the verb labels but can also efficiently localize the action regions in the images. We then explain how we use predicted verbs along with their textual information from sense definitions, object label annotations and image representations extracted from convolutional neural network models for the task of visual sense disambiguation. We describe our proposed unsupervised algorithm based on Lesk which performs visual sense disambiguation using textual, visual and multimodal information. We also verify our findings by using the textual and multimodal embeddings as features in a supervised setting and analyse the performance of visual sense disambiguation task.

Chapter 3 presents our analysis of human eye fixations for the task of assigning verb labels to images. We ask whether the verb prediction model introduced in Chapter 2 correspond to human intuitions about visual verbs or actions. We study whether image regions a verb prediction model identifies as salient for a given verb correlate with the regions fixated by human observers performing a verb classification task. Additionally, we also compare the correlation of human fixations against visual saliency models and center bias models.

Chapter 4 presents our annotation procedure to create our MultiSense crosslingual visual sense disambiguation dataset. We propose cross-lingual visual sense disambiguation models with different ways of combining textual and visual features and show that multimodal models perform best for the cross-lingual visual sense disambiguation task. We also demonstrate that visual sense disambiguation output can be plugged into standard unimodal machine translation systems to improve both verb accuracy in translations as well the quality of the translated sentences.

Chapter 5 is concerned with using visual information for multilingual representation learning. We present two new models hat jointly learn multilingual multimodal representations for mapping images and sentences into common embedding space us-

ing the image as a pivot between languages. Our proposed models map sentences from multiple languages and images into a common space and evaluate the usefulness of the second language in multimodal search with main focus on advancing multilingual versions of image search and image understanding. Our experiments show state-of-the art results on image-description ranking, semantic textual similarity and competitive results on crosslingual image description generation and crosslingual retrieval tasks. Our models suggest that exploiting multilingual and multimodal resources can help in learning better semantic representations that are useful for various multimodal natural language understanding tasks.

Chapter 6 concludes the thesis by summarizing our main findings, discusses the limitations of our work and points out directions we wish to pursue for further research.

Chapter 2

Disambiguating Visual Verbs

Action recognition in still images is the task of identifying whether an action can be perceived in an image or not. During the past decade, computer vision research on action recognition was targeted on constructing machine learning models to label an action in a image which is a verb phrase. Most of these models are learned using large databases of examples labelled manually by humans. This process of annotation is laborious and time consuming. In addition, these datasets and models do not address ambiguity. In this chapter, we introduce the visual verb sense disambiguation task: given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image. We disambiguate the meaning of a verb shown in an image using the existing linguistic sense-inventories which are well known and used in textual word disambiguation tasks. We introduce a new dataset, which we call VerSe (short for **Ver**b **Se**nse) that augments existing multimodal datasets (COCO and TUHOI) with verb and sense labels. In this chapter we test the hypothesis that visual and textual context associated with the image contributes to the verb sense disambiguation task. We explore both supervised and unsupervised models for the sense disambiguation task using textual, visual, and multimodal embeddings. We also consider a scenario in which we must detect the verb depicted in an image prior to predicting its sense i.e., there is no verbal information associated with the image.

2.1 Motivation

Action recognition, the task of identifying the actions depicted in videos or still images, is a widely studied problem in computer vision. Several applications stand to benefit from the ability to recognise actions, such as image description generation, image/video retrieval, surveillance, and a variety of systems involving human-computer interaction. The bulk of existing work has focused on video data, where motion and temporal information provide cues for recognizing actions. The absence of such cues renders the task more challenging in still images for differentiating actions such as *walking* vs *running* or *opening door* vs *closing door*. Nevertheless, attempts to recognise actions in images can be broadly grouped into:

(a) Action Classification (AC), which aims to label an image with a verb phrase, typically a combination of a verb and its object (e.g., *play baseball*, *ride horse*), while assuming that such labels are mutually exclusive (Ikizler et al., 2008; Gupta et al., 2009; Yao & Fei-Fei, 2010; Everingham et al., 2010; Yao et al., 2011).

(b) Human Object Interaction (HOI) recognition, which aims to identify all possible interactions between a human and an object in an image; co-occurring actions (e.g., *hold bicycle* and *ride bicycle*) can in principle be modeled since images receive multiple labels (Le et al., 2014; Chao et al., 2015a; Lu et al., 2016a).

(c) Visual Semantic Role Labeling (VSRL), which identifies the roles actors and objects play in the activity or situation depicted in the image (Gupta & Malik, 2015; Yatskar et al., 2016).

In Figure 2.1 we illustrate each of these tasks and how they relate to each other.

However, none of these action recognition tasks considers the ambiguity that arises when verbs are used as labels. For example, the verb *play* has multiple meanings in different contexts: participate in sport, play musical instrument, or engage in playful activity. ImSitu dataset proposed by Yatskar et al. (2016) explicitly avoided polysemous verbs such as *play*. Moreover, action labels consisting of verb-object pairs may miss important generalizations, e.g., the fact that *ride horse* and *ride elephant* both evoke the same verb semantics, namely *ride animal*. Existing action labels also miss generalizations across verbs, e.g., the fact that *fix bike* and *repair bike* are semantically equivalent, in spite of the use of different verbs. These observations strongly suggest that actions should be analyzed at the level of *verb senses,* similarly to how they are studied in natural language processing.

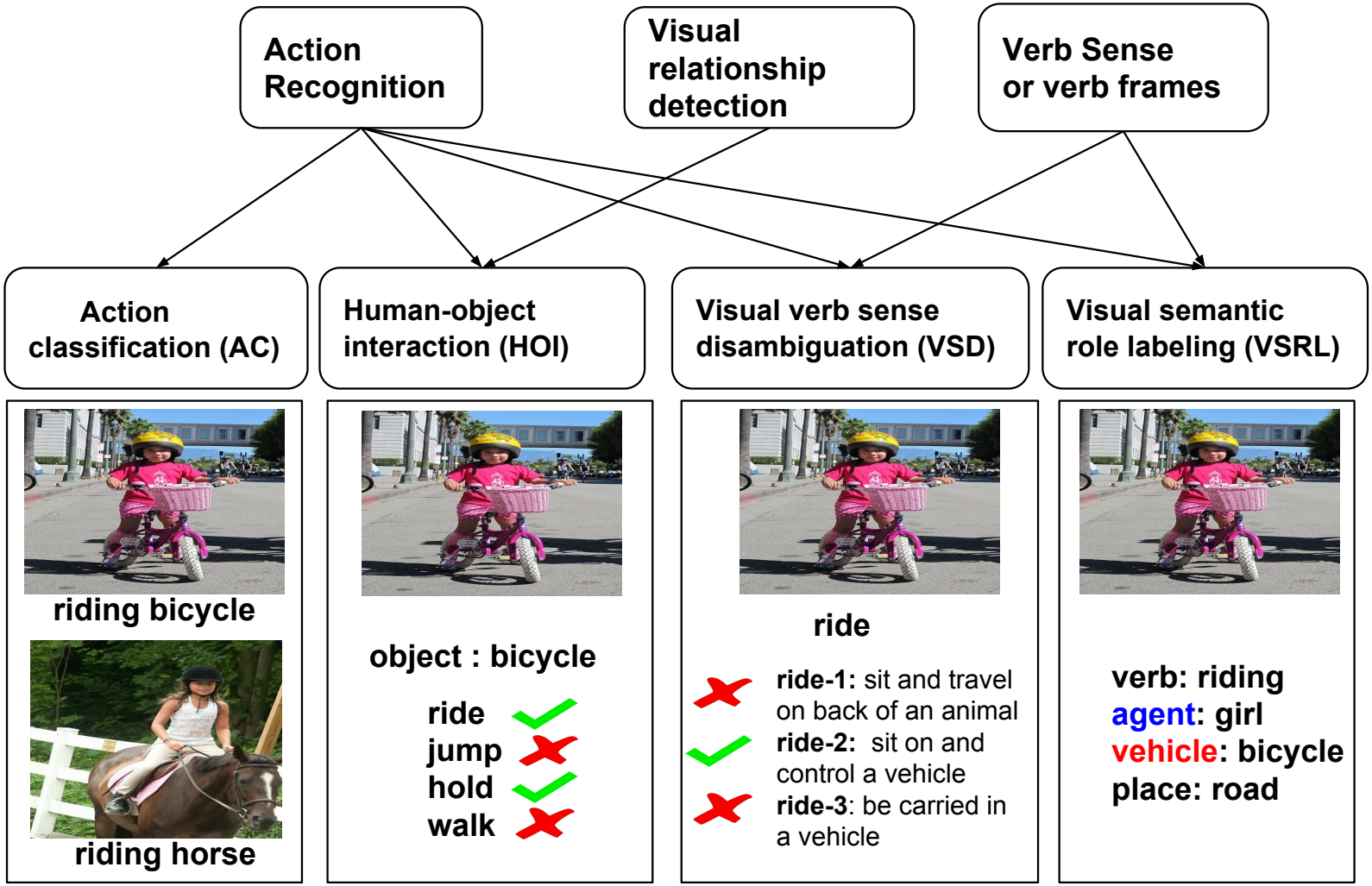


Figure 2.1: Categorization of action recognition tasks in images.

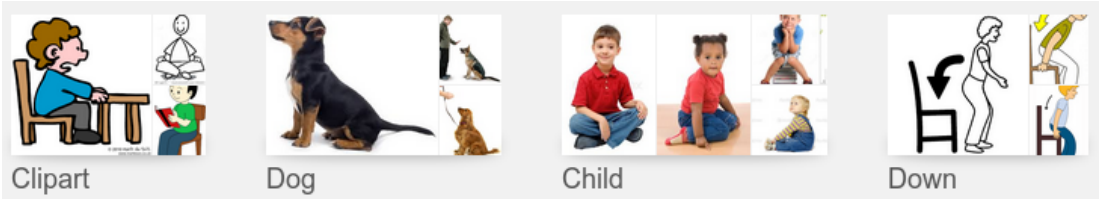


Figure 2.2: Google Image Search trying to disambiguate *sit*. All clusters pertain to the sit down sense with different objects involved, other senses such as baby sitor convene are not included.

In this chapter, we therefore propose the new task of visual verb sense disambiguation (VSD), which aims to label an image with a verb sense taken from a lexical database (see Figure 2.1). We explore two VSD scenarios: (1) given an image and a verb, assign the correct sense of the verb, i.e., the one that describes the action depicted in the image; and (2) given an image, predict a verb and its corresponding sense to correctly describe the action in the image. We present VerSe, a new dataset that augments existing multimodal datasets (COCO and TUHOI) with sense labels. VerSe contains 3,510 images, each annotated with one of 90 verbs, as well as the verb sense realized in the image according to the OntoNotes sense inventory (Hovy et al., 2006).

For our first scenario, we explore both unsupervised and supervised disambiguation methods. Note that our unsupervised methods are unsupervised for sense disambiguation task. However, they do leverage prior supervision from other models that were used to learn objects or generate captions.

We focus in particular on how to best represent word senses for visual disambiguation, and explore the use of textual, visual, and multimodal embeddings. Textual embeddings for a given image can be constructed over object labels or image descriptions, which are available as gold-standard in the COCO and TUHOI datasets, or can be computed automatically using object detectors and image description models. We have experimented with both gold-standard and predicted object labels and image descriptions. Our results show that textual embeddings perform best when gold-standard textual annotations are available, while multimodal embeddings perform best when automatically generated object labels are used. Interestingly, we find that automatically generated image descriptions result in inferior performance. For our second scenario, we predict the verbs depicted in an image using multilabel classification algorithms, which can operate on bounding boxes from an image or on the full image. Our results show that multiple instance learning (MIL), which takes inputs of positive and negative

bounding boxes for every label, performs better than a multilabel CNN architecture.

In the remainder of this chapter, we first present an overview of related work. We then introduce the VerSe dataset and describe our annotation procedure. Next, we provide the details of our disambiguation and verb prediction models. Experimental results and discussion conclude the chapter.

2.2 Related Work

Sense Disambiguation Visual sense disambiguation is related to word sense disambiguation (WSD), a canonical task in natural language processing. The aim in WSD is to identify the intended meaning (sense) of a word in its *textual context.* For example consider the following sentences:

- A man is *playing* a fiddle in the living room.
- The children are *playing* across the street.
- Two men are *playing* doubles tennis on a grass court.

The respective occurrence of play are used with different meaning. The first occurrence corresponds to a "perform or transmit music", second sentence correspond to "engage in a fun or recreational (childlike) activity" and the third one corresponds to a "engage in or make moves related to competition or sport". It is obvious in most cases for humans to interpret this difference whereas it is a difficult task to computationally distinguish between the two different meanings of the word. Reliable WSD has been argued to improve a range of NLP applications, including information retrieval, information extraction, machine translation, content analysis, and lexicography.

There is an extensive literature on WSD for nouns, verbs, adjectives, and adverbs. Most of these approaches rely on lexical databases and sense inventories such as WordNet (Miller et al., 1990) or OntoNotes (Hovy et al., 2006). Unsupervised WSD approaches often rely on distributional representations, computed over the target word and its context (Lin, 1997; McCarthy et al., 2004; Brody & Lapata, 2008; Navigli, 2009). Most supervised approaches use sense annotated corpora to extract linguistic features of the target word (context words, part-of-speech tags, collocation features), which are then fed into a classifier to disambiguate test data (Zhong & Ng, 2010). Recently, features based on sense-specific semantic vectors learned using large corpora and a sense inventory have been shown to achieve state-of-the-art results for supervised WSD (Rothe & Schutze, 2015; Jauhar et al., 2015).

In a multimodal setting (e.g., newspaper articles with photographs), *visual context* is also available and can be used for sense disambiguation in multimodal tasks such as image retrieval. As an example, consider the output of Google Image Search for the query *sit*, shown in Figure 2.2: the search engine recognises that the verb has multiple senses and tries to cluster relevant images. However, the result does not capture the polysemy of the verb well, and would clearly benefit from visual sense disambiguation for clustering images according to its meaning.

In the literature, VSD has been attempted only for nouns (e.g., *apple* can mean fruit or computer. Sense discrimination for web images was introduced in (Loeff et al., 2006), who used spectral clustering over multimodal features from images and web text. (Saenko & Darrell, 2008) employ sense definitions from a dictionary to learn a latent LDA space overs senses, which is then used to construct sense-specific classifiers by exploiting the text surrounding an image.

In general, VSD for nouns is a relatively straightforward task that can be solved with the help of an object detector (Barnard et al., 2003; Chen et al., 2015b). This is helped by resources such as ImageNet (Deng et al., 2009), a large image database containing 1.4 million images for 21,841 noun senses and organised according to the WordNet hierarchy. However, we are not aware of any previous work on VSD for verbs, and no ImageNet for verbs exists. Not only image retrieval would benefit from VSD, but also other multimodal tasks that have recently received a lot of interest, such as automatic image description (Bernardi et al., 2016) and visual question answering (Antol et al., 2015), multimodal machine translation (Elliott et al., 2016; Specia et al., 2016).

Action Recognition As mentioned in Section 3.1, our work relates to a variety of action recognition tasks. To elucidate key aspects of VSD and differences from previous approaches, we provide an overview of commonly used datasets for action recognition in Table 2.1. We observe that the number of verbs covered in these datasets is often smaller than the number of action labels reported (see columns #V and #L) and in many cases the action labels involve an object reference. A few of the first action recognition datasets (e.g., Ikizler (Ikizler et al., 2008) and Willow (Delaitre et al., 2010)) were taken from the sports domain, aiming to capture variation in human poses for actions such as *tennis serve* and *cricket bowling*. As a result, they contain images exhibiting diversity in camera view point, background, and resolution. Further datasets were created based on the intuition that object information helps in modeling action recognition (Li & Fei-Fei, 2007; Ikizler-Cinbis & Sclaroff, 2010), using mutually exclusive labels such as *ride*

horse or *ride bike*.

The limitations of the early datasets (small size, domain specificity, and the use of ad-hoc labels) have been recently addressed in a number of broad-coverage resources that are large scale and use linguistically-motivated labels (Yatskar et al., 2016; Ronchi & Perona, 2015; Chao et al., 2015a). Often these datasets use existing linguistic resources such as VerbNet (Schuler, 2005), WordNet, and FrameNet (Baker et al., 1998) to classify verbs. This allows for a more general, semantically motivated treatment of verbs and verb phrases, and also takes into account the fact that not all verbs are depictable. For example, abstract verbs such as *presume* and *acquire* are not depictable, while other verbs have both depictable and non-depictable senses: *play* is non-depictable in *play with emotions*, but depictable in *play an instrument* and *play a sport*. A few other datasets have been based on Microsoft Common Objects in Context (COCO; (Chen et al., 2015a)), a dataset that consists of over 120k images with extensive annotations, including labels for 91 object categories and five descriptions per image. Although COCO was not created with action recognition in mind, it is possible to use the verbs present in the descriptions to annotate actions and their semantic roles (Ronchi & Perona, 2015; Gupta & Malik, 2015).

It is important to note that verb sense ambiguity is ignored in almost all existing action recognition datasets (and corresponding tasks). This misses important generalizations: for instance, the actions *ride horse* and *ride elephant* represent the same sense of *ride* and thus share visual, textual, and conceptual features. On the other hand, *play tennis* and *play guitar* share the same verb but represent different senses. We address this issue by creating VerSe, a dataset with explicit sense labels. VerSe is built on top of TUHOI (the Trento Universal Human-Object Interaction dataset; (Le et al., 2014)) and COCO. The former dataset contains 10,805 images covering 2,974 actions. Action categories were crowdsourced, each image was labeled by multiple annotators with a description in the form of a verb or a verb-object pair. The main drawback of TUHOI is that 1,576 out of 2,974 action categories occur only once, limiting its usefulness for VSD. Although COCO contains no explicit action annotation, verbs and verb phrases can be extracted from the descriptions. (But note that only about half of the COCO images depict actions.)

The HICO (Humans Interacting with Common Objects) dataset is conceptually similar to VerSe. It consists of 47,774 images annotated with 111 verbs and 600 human-object interaction categories.

Dataset	Task	#L	#V	Obj	Images	Sen	Des	Cln	ML	Resource	Example Labels
Ikizler (Ikizler et al., 2008)	AC	6	6	0	467	N	N	Y	N	—	run, walk
Sports Dataset (Gupta et al., 2009)	AC	6	6	4	300	N	N	Y	N	—	tennis serve, cricket bowling
Willow (Delaitre et al., 2010)	AC	7	6	5	986	N	N	Y	Y	—	ride bike, take photograph
PPMI (Yao & Fei-Fei, 2010)	AC	24	2	12	4.8k	N	N	Y	N	—	play guitar, hold violin
Stanford 40 Actions (Yao et al., 2011)	AC	40	33	31	9.5k	N	N	Y	N	—	cut vegetables, ride horse
PASCAL 2012 (Everingham et al., 2015)	AC	11	9	6	4.5k	N	N	Y	Y	—	ride bike, ride horse
89 Actions (Le et al., 2013a)	AC	89	36	19	2k	N	N	Y	N	—	ride bike, fix bike
MPII Human Pose (Andriluka et al., 2014)	AC	410	—	66	40.5k	N	N	Y	N	—	ride car, hair styling
TUHOI (Le et al., 2014)	HOI	2974	—	189	10.8k	N	N	Y	Y	—	sit on chair, play with dog
BU101 Dataset (Ma et al., 2017)	AC	101	68	—	23.8k	N	N	Y	N	—	horse race, play violin
COCO-a (Ronchi & Perona, 2015)	HOI	—	140	80	10k	N	Y	Y	Y	VerbNet	walk bike, hold bike
Google Images (Ramanathan et al., 2015)	AC	2880	—	—	102k	N	N	N	N	—	riding horse, riding camel
HICO (Chao et al., 2015a)	HOI	600	111	80	47k	Y	N	Y	Y	WordNet	ride#v#1 bike; hold#v#2 bike
VCOCO-SRL (Gupta & Malik, 2015)	VSRL	—	26	48	10k	N	Y	Y	Y	—	verb: hit; instrument: bat; object: ball
imSitu (Yatskar et al., 2016)	VSRL	—	504	11k	126k	Y	N	Y	N	FrameNet WordNet	verb: ride; agent: girl#n#2 vehicle: bike#n#1; place: road#n#2
VerSe (Ours)	VSD	163	90	—	3.5k	Y	Y	Y	N	OntoNotes	ride.v.01, play.v.02

Table 2.1: Comparison of existing action recognition datasets according to various subtasks. #L denotes the number of action labels in the dataset; #V denotes the number of verbs covered in the dataset; Obj indicates the number of objects annotated; Sen indicates whether sense ambiguity is explicitly handled; Des indicates whether image descriptions are included; Cln denotes whether the dataset has been manually verified; ML indicates the possibility of multiple labels per image

Unlike other existing datasets, HICO uses sense-based distinctions: actions are denoted by sense-object pairs, rather than by verb-object pairs. HICO does not aim for complete coverage of senses: it restricts itself to a single sense of a verb (with the exceptions of a couple of verbs), which means that HICO is not suitable for verb sense disambiguation.

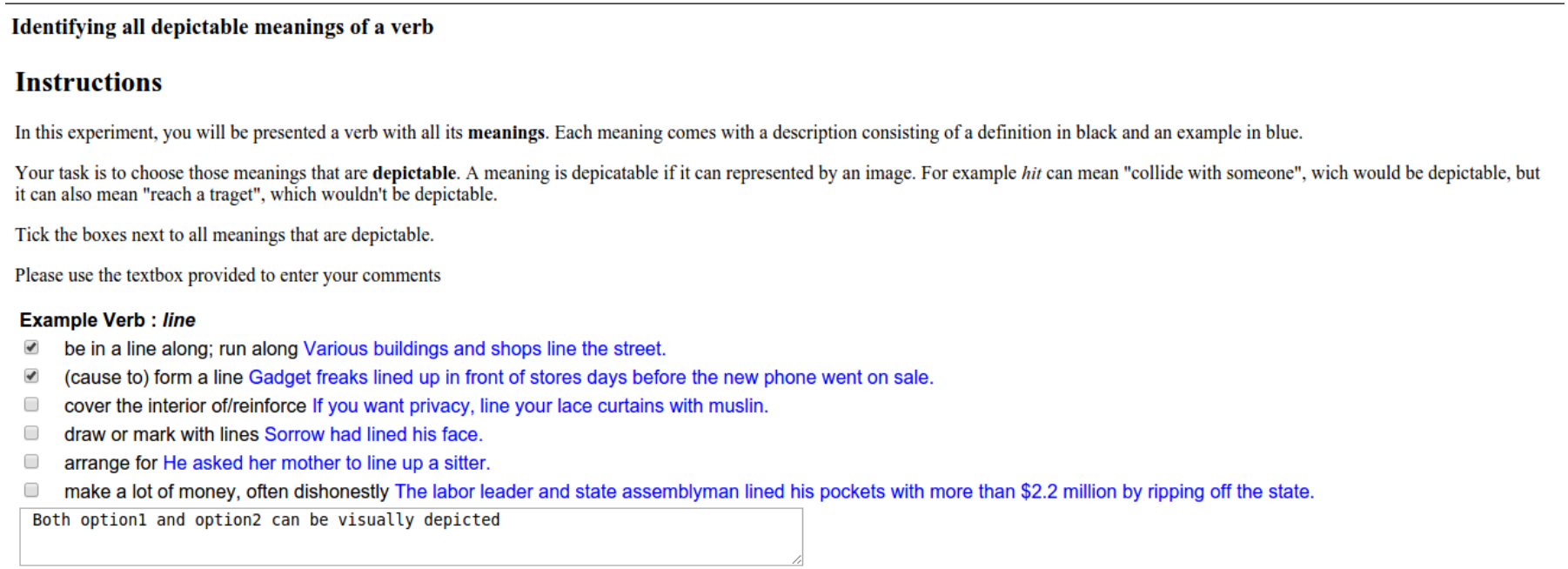
The COCO-a dataset (Ronchi & Perona, 2015) was created by identifying verbs that are visual and detectable in images. The selection criteria included that a 6–8 year old child should be able to distinguish the visual verbs. This strategy meant that synonyms or related verbs were not included in the dataset, and also polysemous uses of verbs were excluded. The authors cross-checked the verbs they selected against the verbs used in the COCO image descriptions. This resulted in a total of 140 visual verbs being covered in COCO-a.

Another dataset is imSitu (Yatskar et al., 2016), which includes a large number of images and annotates each image with a verb and its semantic frames taken from FrameNet (Baker et al., 1998). Each semantic frame includes a frame label (e.g., gardening), the frame elements (e.g., agent, tool), and the location (e.g., outdoors). The frame annotation by definition determines the sense of a verb. However, when imSitu was designed, it was decided not to include polysemous verbs, so for example the verb *play* is not in the dataset. Because all the verbs in the dataset only have one sense, imSitu cannot be used for visual sense disambiguation.

Our VerSe dataset presented in Section 2.3 addresses the visual sense disambiguation problem. The annotation of VerSe proceeds into two steps: 1) Identifying the verbs that are visual (similar to the work of (Chao et al., 2015b; Ronchi & Perona, 2015)) 2) Identifying all the visual senses of a verb and annotating images with verb visual senses. To identify the target set of verbs we sampled the verbs from existing datasets TUHOI and COCO.

2.3 The VerSe Dataset

In this section we describe how VerSe was created. As mentioned earlier, it is based on previous image description and action recognition datasets COCO and TUHOI, covering 90 verbs, and contains 3,518 images. VerSe serves two main purposes: (1) to show the feasibility of annotating images with verb senses (rather than verbs or actions); (2) to function as test bed for evaluating automatic visual sense disambiguation methods.



Identifying all depictable meanings of a verb

Instructions

In this experiment, you will be presented a verb with all its **meanings**. Each meaning comes with a description consisting of a definition in black and an example in blue.

Your task is to choose those meanings that are **depictable**. A meaning is depicatable if it can represented by an image. For example *hit* can mean "collide with someone", wich would be depictable, but it can also mean "reach a traget", which wouldn't be depictable.

Tick the boxes next to all meanings that are depictable.

Please use the textbox provided to enter your comments

Example Verb : *line*

- ☑ be in a line along; run along Various buildings and shops line the street.
- ☑ (cause to) form a line Gadget freaks lined up in front of stores days before the new phone went on sale.
- ☐ cover the interior of/reinforce If you want privacy, line your lace curtains with muslin.
- ☐ draw or mark with lines Sorrow had lined his face.
- ☐ arrange for He asked her mother to line up a sitter.
- ☐ make a lot of money, often dishonestly The labor leader and state assemblyman lined his pockets with more than $2.2 million by ripping off the state.

Both option1 and option2 can be visually depicted

Figure 2.3: Example item for depictability and sense annotation: sense definitions and examples (in blue) for the verb *line*.

Verb Selection Action recognition datasets often use a limited number of verbs in a given domain (see Table 2.1). We instead sampled verbs from COCO descriptions and TUHOI verb phrases (e.g., *sit on chair*), which we use in lieu of descriptions. We extracted all verbs from all descriptions in the two datasets and selected those with more than one sense in the OntoNotes dictionary (Hovy et al., 2006). This procedure resulted in 148 verbs in total (94 from COCO and 133 from TUHOI).

Depictability Annotation A verb can have multiple senses, but not all of them are depictable, e.g., senses describing cognitive and perception processes are not depictable. Consider the verb *touch* whose make physical contact sense is depictable, whereas the affect emotionally sense is not depictable. We therefore annotated the senses of a verb as depictable or non-depictable. Amazon Mechanical Turk (AMT) workers were presented with the definitions of all the senses of a verb, along with examples, as given by OntoNotes (Hovy et al., 2006). An example for this annotation is shown in Figure 2.3. We used OntoNotes instead of WordNet, as WordNet senses are very fine-grained and potentially make depictability and sense annotation harder. Granularity issues with WordNet for text-based WSD are well documented (Navigli, 2009).

OntoNotes (Hovy et al., 2006), a coarse grained sense inventory is created to address the sense granularity issue of WordNet, by iteratively partitioning WordNet senses until they reach and inter-annotator agreement of 90% on the sense annotation task. OntoNotes senses can be mapped to their respective WordNet senses, for verb senses they can also be mapped Propbank roles and FrameNet frames.

Although on the surface our depictability labels might look similar to affordance visualness labels (Chao et al., 2015b), there are two main key differences: 1) Their annotation is restricted to few senses per each verb, whereas we do exhaustive annotation of all senses in the verb which provides an overview of visualness of not just the synset but the overall verb. They concentrated on covering a higher number of verbs with fewer senses, whereas we covered fewer verbs than theirs but all of the verb senses. 2) They use a rating scale of $1-5$, where score 1 represents "definitely not visual or makes no sense" to score 5 which represents "definitely visual" where as we use much simpler binary labels of yes/no to determine the visualness of the sense.

OntoNotes lists 921 senses for our 148 target verbs. For each verb, three AMT workers selected all depictable senses. The majority label was used as the gold-standard for subsequent experiments. This resulted in a 504 depictable senses. Inter-annotator agreement (ITA) as measured by Fleiss' Kappa was 0.645. Annotated visual senses for verbs *play* and *serve* are listed in Appendix A.1.

Identifying all depictable meanings of a verb

Instructions

In this experiment, you will be presented two image, verb pairs with a set of verb **meanings**. Each meaning comes with a description consisting of a definition in black and examples in blue.

Your task is to choose those meaning of the verb that is depictable from the given image. If none of the provided meanings is applicable, choose **None of the above** option provided.

If you have any comments, please use the textbox provided.

Example

Verb: hit

- strike with an instrument, missile, or oneself Jasmine hit the ball to her cousin. more examples
- consume to excess Every time she gets upset, she hits the bottle. more examples
- (•) perform a type of ball striking in a sport Williams hit a pop fly to center field. more examples
- pay unsolicited sexual attention to He likes to go to bars and hit on women.
- throw oneself to the ground When the shots rang out, everyone hit the dirt.
- go to sleep I need to hit the hay. more examples
- None of the above

Option3 is selected as this picture depicts a woman hitting a ball

Figure 2.4: Annotation guidelines for verb sense annotation for the given image and verb: *hit*

Verb type	Examples	Verbs	Images	Senses	Depct	ITA
Motion	run, walk, jump, swing, hit, kick, etc.	39	1812	10.76	5.79	0.680
Non-motion	sit, sleep, lean, read, stand, lay, etc.	51	1698	8.27	4.86	0.636

Table 2.2: Overview of VerSe dataset divided into motion and non-motion verbs; Depct: depictable senses; ITA: inter-annotator agreement.

Sense Annotation We then annotated a subset of the images in COCO and TUHOI with verb senses. An image was assigned the verb that occurs most frequently in the descriptions for that image (for TUHOI, the descriptions are verb-object pairs, see above). Although multiple verbs can be applicable in a given image, we only annotated the most frequently occurring verb. Perhaps not surprisingly, we observed that the distribution of verbs and their corresponding images is Zipfian: there are many verbs represented by a few images, and a few verbs represented by a large number of images. For sense annotation, we selected only verbs for which either COCO or TUHOI contained five or more images, resulting in a set of 90 verbs (out of the total 148). An example of our sense annotation is shown in Figure 2.4. All images for these verbs were included, resulting in a dataset of 3,528 images: 2,340 images for 82 verbs from COCO and 1,188 images for 61 verbs from TUHOI (some verbs occur in both datasets).

These image-verb pairs formed the basis for sense annotation. AMT workers were presented with the image and all the depictable OntoNotes senses of the associated verb. The workers had to chose the sense of the verb that was instantiated in the image (or "none of the above", in the case of irrelevant images). Annotators were given sense definitions and examples, as in the depictability annotation (see Figure 2.3). For every image-verb pair, five annotators performed the sense annotation task. A total of 157 annotators participated, reaching an inter-annotator agreement of 0.659 (Fleiss' Kappa). Out of 3,528 images, we discarded 18 images annotated with "none of the above", resulting in a set of 3,510 images covering 90 verbs and 163 senses. Number of images per verb sense varied from $1-100$. We present statistics of our dataset in Table 2.2; we group the verbs into motion verbs and non-motion verb using Levin verb classes (Levin, 1993). An example of annotated images group according to senses for verbs *play* and *serve* are listed in Appendix A.2.

Figure 2.5: Visual sense ambiguity: three of the senses of the verb *play*: play sport, play instrument, children play.

2.4 Visual Verb Sense Disambiguation

For our disambiguation task, we assume we have a set of images I, and a set of polysemous verbs V and each image $i \in I$ is paired with a verb $v \in V$. For example, Figure 2.5 shows different images paired with the verb *play*. Every verb $v \in V$, has a set of senses $\mathcal{S}(v)$, described in a dictionary $\mathcal{D}$. Now, given an image i paired with a verb v, our task is to predict the correct sense $\hat{s} \in \mathcal{S}(v)$, i.e., the sense that is depicted by the associated image. In Figure 2.5, the correct sense for the first image is participate in sport, for the second one it is play an instrument, and so on.

The disambiguation task can be performed in a supervised manner, using samples of images, verbs, and their manually annotated senses. In this case, a classifier is used to assign each verb its appropriate sense based on evidence from contextual features extracted from the accompanying image or any textual information available. While this approach often achieves high accuracy, adequately large sense labeled data sets are difficult to obtain across languages and sense inventories. We therefore also explore an unsupervised approach which requires no sense annotated training data. Note, our unsupervised experiments are close to semi-supervised since our models does leverage prior supervision that was used to learn the CNN/captioning/object list. Our experiments do not use VerSe dataset for training and they are only used for evaluating our models. Our denotion of unsupervised refers For unsupervised sense disambiguation, we propose a new variant of the Lesk algorithm (Lesk, 1986), a well-known approach to text-based WSD, which relies on the calculation of the word overlap between the sense definitions and the context in which a word occurs. The algorithm uses the following scoring function to disambiguate the sense of a verb v:

$$\hat{s} = \underset{s \in \mathcal{S}(v)}{\arg\max}\, \Phi(s, v, \mathcal{D}) = |\text{context}(v) \cap \text{definition}(s, \mathcal{D})| \tag{2.1}$$

Here, context(v) is the set of words that occur close to the target word v and definition$(s, \mathcal{D})$ is the set of words in the definition of sense s in dictionary $\mathcal{D}$.

In our case, context(v) is the image i associated with v. We create a representation for a given image (the vector $\mathbf{i}$), which can be text-based (using the object labels and descriptions for i), visual, or multimodal. Similarly, we create text-based, visual, and multimodal representations (the vector $\mathbf{s}$) for every sense s of a verb. Based on the representations $\mathbf{i}$ and $\mathbf{s}$ (detailed below), we score senses as: Taking the dot product of two normalized vectors is equivalent to using cosine as similarity measure. We experimented with other similarity measures, but cosine performed best.

$$\hat{s} = \arg\max_{s \in \mathcal{S}(v)} \Phi(s, v, i, \mathcal{D}) = \mathbf{i} \cdot \mathbf{s} \tag{2.2}$$

An overview of our method is given in Figure 2.6. The various image representations (visual, textual, and multimodal) also serve as features in the supervised setting. In that setting, there is no need to represent senses; the sense are simply labels the classifier learns to predict. In the following, we will describe in more detail how we obtain image and sense representations.

2.4.1 Image Representations

Visual Modality: Creating a visual representation $\mathbf{i}^{\mathbf{c}}$ of an image i is straightforward. We used the VGG 16-layer architecture (VGGNet) trained on 1.2M images of the 1,000 class ILSVRC 2012 object classification dataset, a subset of ImageNet (Simonyan & Zisserman, 2014). This CNN model has a top-5 classification error of 7.4% on ILSVRC 2012. We used the publicly available reference model implemented using CAFFE (Jia et al., 2014) to extract the output of the fc7 layer, i.e., a 4,096 dimensional vector, for every image i. We use this vector as our image representation.

Textual Modality: We also explore the possibility of representing the image indirectly, viz., through text associated with it in the form of object labels (O) or image descriptions (C), as shown in Figure 2.6. We experiment with two different forms of textual annotation: gold-standard (GOLD) annotation, where object labels and descriptions are provided by human annotators, and predicted (PRED) annotation, where state-of-the-art object recognition and image description generation systems are applied to the image.

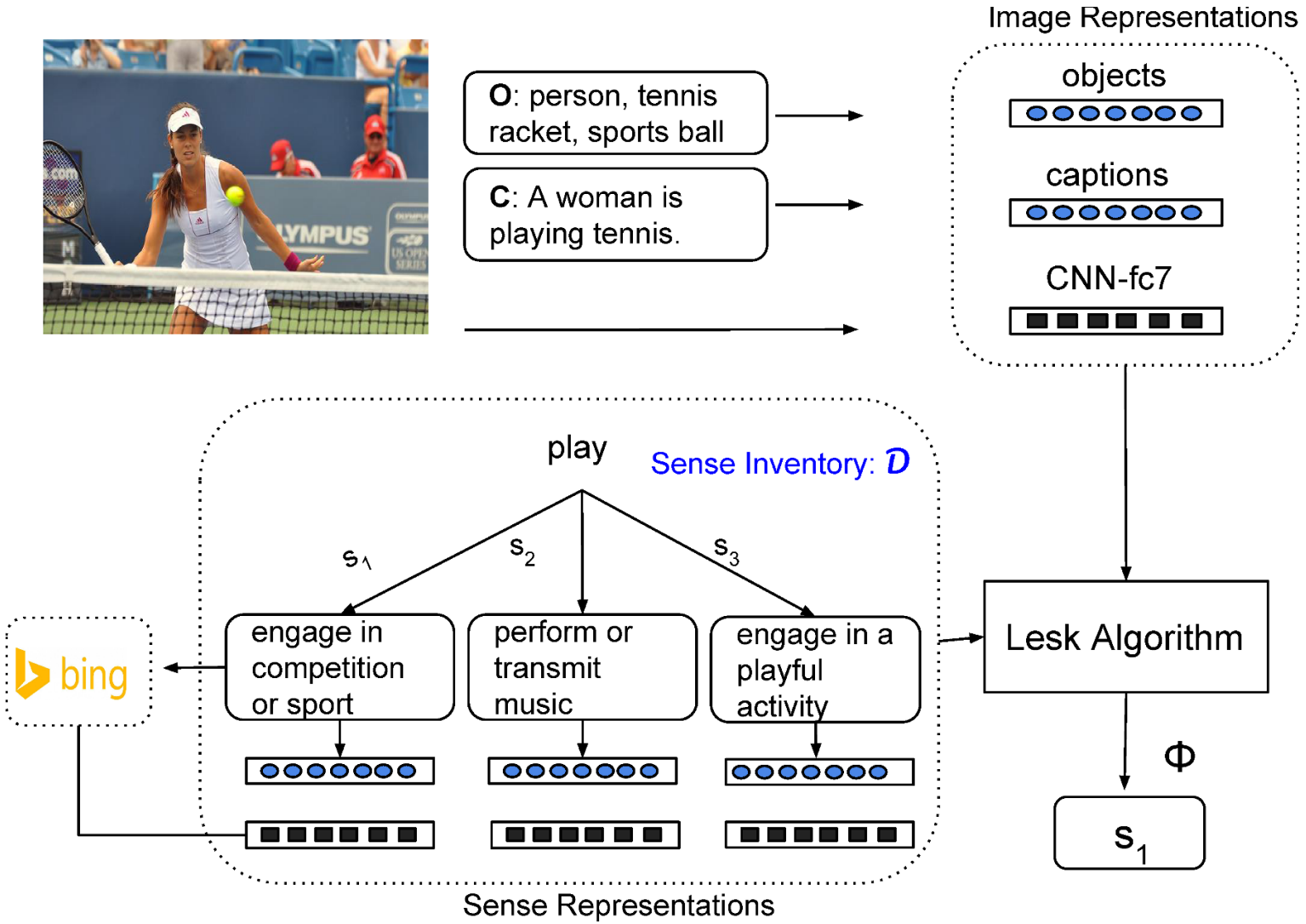


Figure 2.6: Schematic overview of the visual sense disambiguation model.

GOLD object annotations are provided with the two datasets we use. Images sampled from COCO are annotated with one or more of 91 object categories. Images from TUHOI are annotated with one more of 189 object categories. PRED object annotations were generated using the same VGG 16-layer CNN object recognition model that was used to compute visual representations. Only object labels with an object detection threshold $t > 0.2$ were used.

To obtain GOLD image descriptions, we used the used human-generated descriptions that come with COCO. For TUHOI images, we generated descriptions of the form subject-verb-object tuples, where the subject is always *person*, and the verb-object pairs are the action labels that come with TUHOI. To obtain PRED descriptions, we generated three descriptions for every image using the state-of-the-art image description system of Vinyals et al. (Vinyals et al., 2015).We used Karpathy's implementation, publicly available at https://github.com/karpathy/neuraltalk.

We create a textual representation $\mathbf{i^t}$ of image i using word2vec (Mikolov et al., 2013a), a widely used model of word embeddings. Specifically, we obtain a vector for each object label and word in the image descriptions. An overall representation of the image is then computed by averaging these vectors over all labels, all content words in the description, or both. For our experiments we used the pre-trained 300 dimensional vectors available with the word2vec package (trained on part of the Google News dataset, about 100 billion words).

Modality Combination Apart from experimenting with separate textual and visual representations of images, it also makes sense to combine the two modalities into a multimodal representation. The simplest approach is a concatenation model which appends textual and visual features. More complex multimodal vectors can be created using methods such as Canonical Correlation Analysis (CCA; (Hardoon et al., 2004)) and Deep Canonical Correlation Analysis (DCCA; (Andrew et al., 2013b; Wang et al., 2015)). CCA allows us to find a latent space in which the linear projections of text and image vectors are maximally correlated (Gong et al., 2014b; Hodosh et al., 2015). DCCA can be seen as a non-linear version of CCA and has been successfully applied to the image description task (Yan & Mikolajczyk, 2015a), outperforming previous approaches, including kernel-based CCA.

We employ both CCA and DCCA to map the vectors $\mathbf{i^t}$ and $\mathbf{i^c}$ (which have different dimensions) into a joint latent space of $n = 300$ dimensions. We represent the projected vectors of textual and visual features for image i as $\mathbf{i^{t'}}$ and $\mathbf{i^{c'}}$ and combine them to

obtain a multimodal representation $\mathbf{i^m}$ as follows:

$$\mathbf{i^m} = \lambda \mathbf{i^{t'}} + (1-\lambda)\mathbf{i^{c'}} \tag{2.3}$$

where λ is a parameter representing the relative importance of the textual and visual modalities. We present the details of λ in Section 2.6.1.

2.4.2 Sense Representations

For sense disambiguation, we must also obtain representations for verb senses (see Equation (2.2)). Analogously to image representations, we create a visual sense representation $\mathbf{s^c}$, a text-based sense representation $\mathbf{s^t}$, and one that combines both modalities.

Visual Modality Sense dictionaries typically provide sense definitions and example sentences, but no visual examples or images. For nouns, this is remedied by ImageNet (Deng et al., 2009), which provides a large number of example images for a subset of the senses in the WordNet noun hierarchy. However, no comparable resource is available for verbs (see Section 4.2).

In order to obtain visual sense representation $\mathbf{s^c}$, we therefore collected sense-specific images for the verbs in our dataset. For each verb sense s, three trained annotators were presented with the definition and examples from OntoNotes, and had to formulate a query $Q(s)$ that would retrieve images depicting the verb sense when submitted to a search engine. Our query formulations are retrived from both sense definitions and examples and predominantly contained verb phrases. For every query q we retrieved images $I(q)$ using the Bing image search engine (for examples, see Figure 2.7). We used the top 50 images returned by Bing per query. For every verb sense s on an average we used 5 queries to retrieve images from Bing.

Images were converted into feature representations, using the output of the fc7 layer of VGGNet (same setup as in Section 2.4.1). To generate a visual representation for an individual sense $\mathbf{s^c}$, we perform mean pooling over the images obtained using the sense specific queries:

$$\mathbf{s^c} = \frac{1}{n} \sum_{q_j \in Q(s)} \sum_{i \in I(q_j)} \mathbf{c_i} \tag{2.4}$$

where n is the total number of images retrieved per sense s.

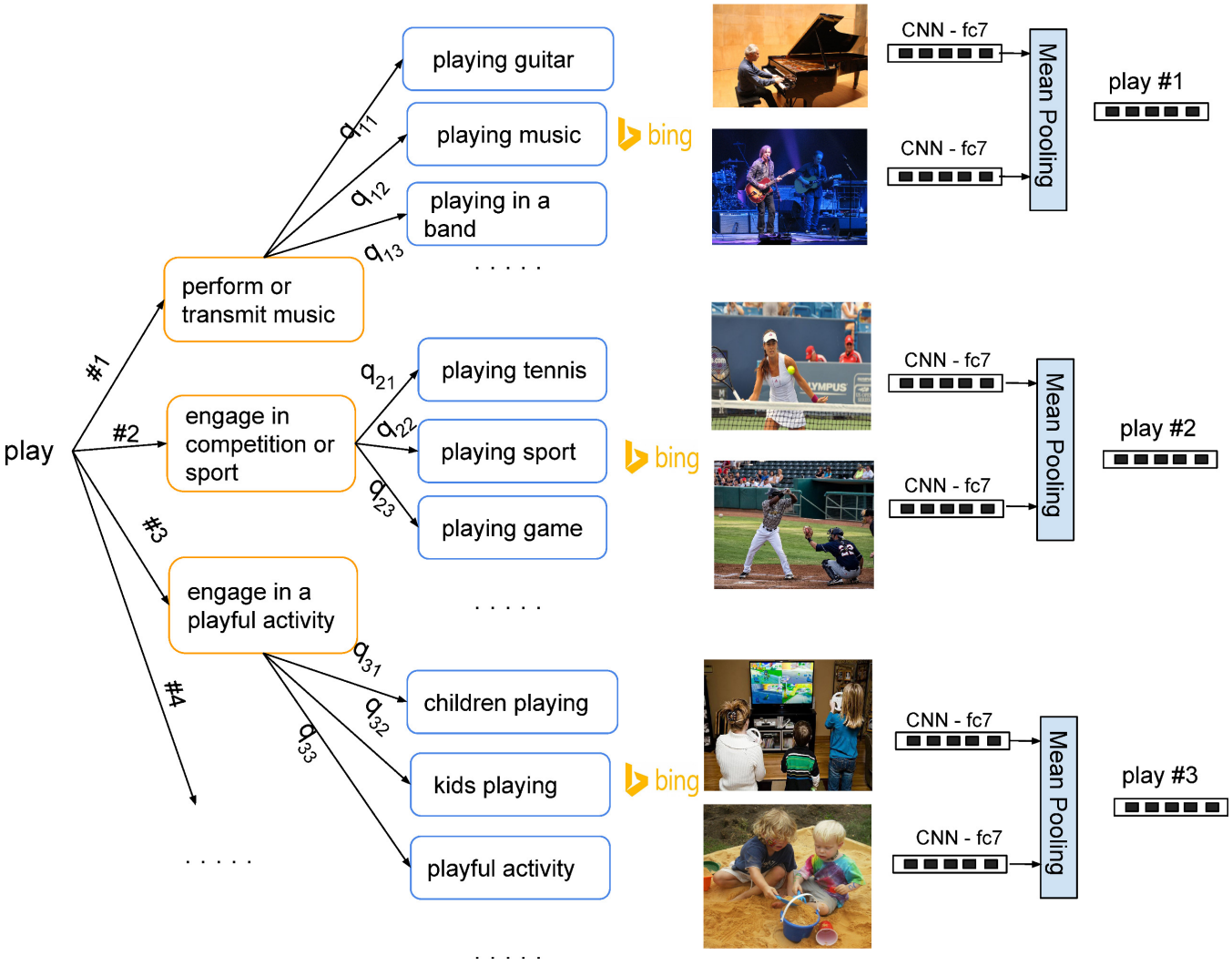


Figure 2.7: Our schematic overview of the visual sense representation for the verb *play* from its sense definitions and examples using Bing Image Search.

Text-based Sense Representation We create a vector $\mathbf{s^t}$ for every sense $s \in \mathcal{S}(v)$ of a verb v from its definition and the example usages provided in the OntoNotes dictionary $\mathcal{D}$. Again, we apply word2vec (Mikolov et al., 2013a) to obtain a vector for every content word in the definition and examples of the sense and take the average of these vectors to compute an overall representation of the verb sense.

Modality Combination Visual and textual modalities for senses were combined as explained previously for images. We obtain a multimodal representation for sense s as follows:

$$\mathbf{s^m} = \lambda \mathbf{s^{t'}} + (1-\lambda)\mathbf{s^{c'}} \tag{2.5}$$

where vectors $\mathbf{s^{t'}}$ and $\mathbf{s^{c'}}$ are projections of the visual and textual representations of sense s onto a joint latent space. We use vectors ($\mathbf{i^t}$, $\mathbf{s^t}$), ($\mathbf{i^c}$, $\mathbf{s^c}$), and ($\mathbf{i^m}$, $\mathbf{s^m}$) as described in Equation (2.2) to perform sense disambiguation.

2.5 Verb Prediction

So far we have focused on disambiguating verbs co-occurring with an image. In cases where images are not associated with textual information, it would be natural to first predict a verb representing the action depicted and then predict the verb sense (using the methods introduced in the previous sections). In the following, we describe two methods for predicting verbs given an image: (1) a multilabel CNN-based classification approach which simultaneously predicts all verbs associated with an image; and (2) a multiple instance learning approach which considers bags of positive and negative bounding boxes to decide which verb is compatible with the image.

2.5.1 Multilabel Classification

We trained a multilabel CNN to simultaneously predict all verbs depictable in a novel test image. Our vocabulary $\mathcal{V}$ consists of the 250 most common verbs (including the 90 verbs in VerSe) in the descriptions of TUHOI, Flickr30k, and COCO datasets. We included Flickr30k as it has a more diverse distribution of verbs compared to COCO and the descriptions are action oriented (Young et al., 2014a). A verb label is considered positive if it appears in the description of the image. In Figure 2.3 we present all the positive verb labels extracted from the image descriptions.

We used a sigmoid cross entropy loss and optimized the ResNet 152-layer CNN architecture. We initialized the network weights with the publicly available CNN pre-

A woman is **playing** tennis on a grass field.

A women is **swinging** a tennis racket.

A woman **playing** tennis.

A large crowd is **watching** a tennis match.

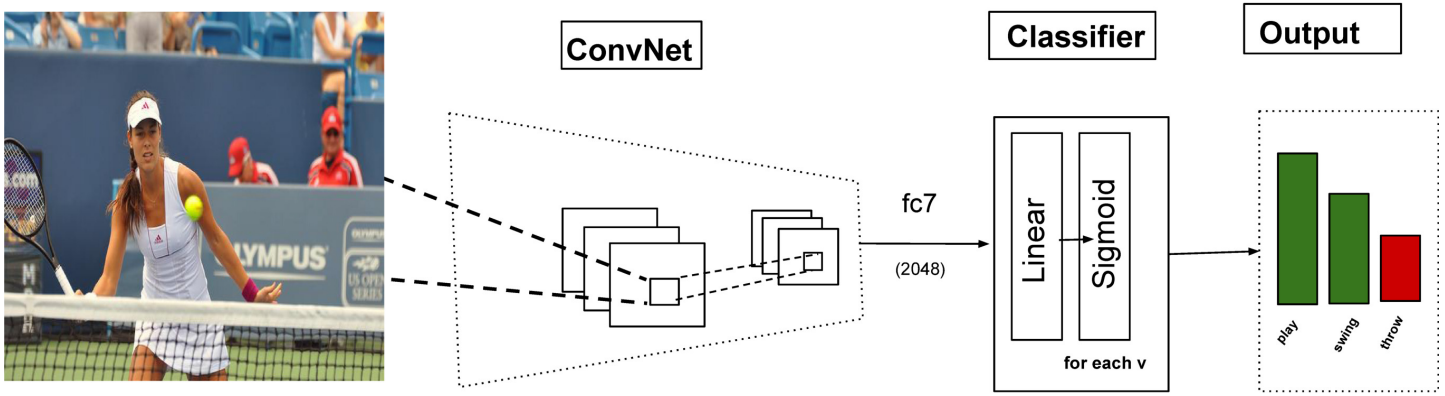
Table 2.3: An example of human annotated descriptions for an image from MSCOCO dataset. We highlight different verbs used in image descriptions.

trained on ImageNet[1] and finetuned it with our own verb labels. We used stochastic gradient descent with momentum set to 0.99 and a learning rate of $1e^{-5}$, i.e., lower than the original network to account for the sparsity of the labels in the training set. The network was trained with a batch size of one for three epochs. The CNN architecture for multilabel classification (MLC) is shown in Figure 2.8(a).

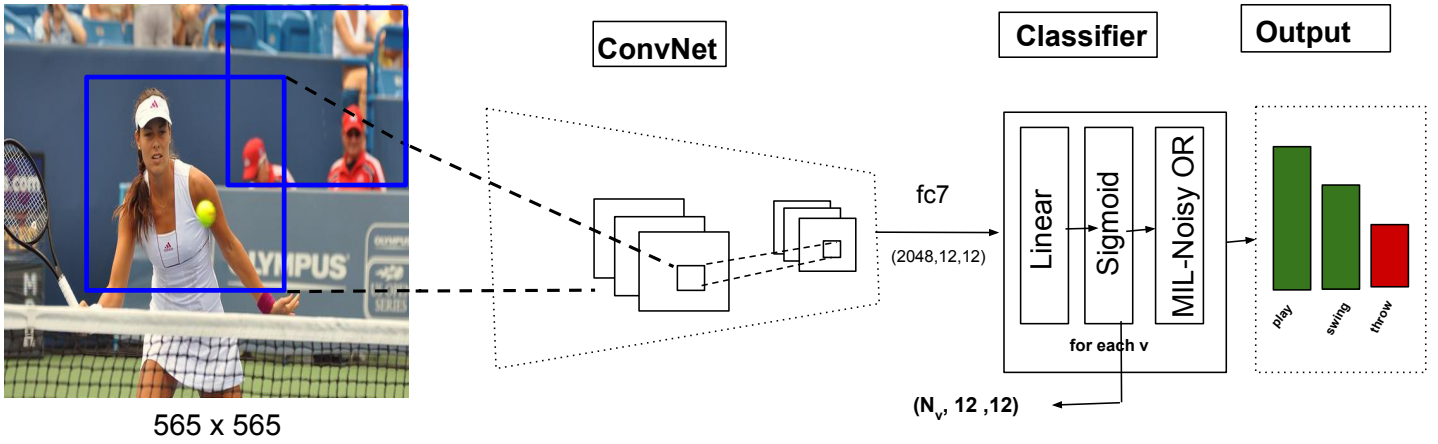
2.5.2 Multiple Instance Learning

In addition to multilabel classification, we experimented with a weakly supervised model based on multiple instance learning (MIL; (Maron & Lozano-Pérez, 1997)) which has shown promising results in a variety of computer vision tasks including object detection (Zhang et al., 2005), image description generation (Fang et al., 2015b), scene classification (Maron & Ratan, 1998), and action recognition (Sener et al., 2012)

[1] https://github.com/KaimingHe/deep-residual-networks#models



(a) Multi-label classification (MLC) model.



(b) Multiple instance classifier (MIL) with Noisy-OR model.

Figure 2.8: Our multi-label verb prediction classifiers highlights the differences between MLC vs. MIL model

For each verb $v \in \mathcal{V}$, MIL samples sets (instances) of "positive" and "negative" bags of bounding boxes, where each bag corresponds to one image i. Each bounding box from the image is drawn based on a sliding window. A bag b_i is positive if verb v is in image i's description, and negative otherwise. During training, instances within the positive bags are iteratively selected and the model is retrained using the updated positive labels. Compared to multilabel classification, which makes predictions considering the image as a whole, MIL is intuitively more appropriate for our task, since different parts of an image could represent different verbs.

We predict p_{ij}^v, the probability that a region j in image i corresponds to verb v, using a multi-layered convolutional neural network architecture which computes a logistic function on top of the last hidden layer (fc7; see (Fang et al., 2015b) for more details):

$$p_{ij}^v = \frac{1}{1+\exp\left(-(\mathbf{w_v}\phi(b_{ij}) + wb_v)\right)} \tag{2.6}$$

where $\phi(b_{ij})$ is the fc7 representation for image region j in image i, and $\mathbf{w_v}$, wb_v are the weights and bias associated with verb v. We then use a noisy-OR version of MIL, where the probability of bag b_i depicting verb v is calculated from the probabilities of the individual instances in the bag:

$$p_i^v = 1 - \prod_{j \in b_i} \left(1 - p_{ij}^v\right) \tag{2.7}$$

Following previous work (Fang et al., 2015b), we upsample images to 565 pixels and use a sliding window of 224×224 with a stride of 32. This results in 144 bounding boxes for each image which we refer to as instances. The noisy-OR version of MIL (Equation (2.7)) is implemented on top of 144 intermediate predictions p_{ij}^v (corresponding to each bounding box region b_{ij}) to compute a single probability p_i^v for each $v \in \mathcal{V}$. We use cross-entropy loss and optimize ResNet-152 (initialized with a CNN network pretrained on ImageNet) end-to-end with stochastic gradient descent[2]. We use the same hyperparameter settings as in multilabel classification for three epochs. At test time, a novel image i is upsampled to 565 pixels to obtain the probability p_i^v for each verb $v \in \mathcal{V}$. The MIL architecture is shown in Figure 2.8(b).

2.6 Experiments

In the following, we report results for two sets of experiments. We first focus on visual sense disambiguation when the input to the system is an image and a verb associated

[2]We also experimented using VGG-19 CNN pre-trained network. However, tey had much lower scores compared to ResNet-152

with it and then move on to the more challenging task of detecting the verbs that are depicted in the image prior to predicting their senses.

2.6.1 Verb Sense Disambiguation

Table 2.4 summarises the results of the unsupervised disambiguation method introduced in Section 2.4. We present results separately for motion and non-motion verbs in our gold-standard (GOLD) and predicted (PRED) settings. As explained earlier, we represent images and their senses by individual modalities (textual or visual) or their combination. To train the CCA and DCCA models, we use the text representations learned from image descriptions in the COCO and Flickr30k datasets as one view and the VGG-16 features from the respective images as the second view. We divide the data into train, test and development samples (using an 80/10/10 split). We use the trained models to generate the projected representations of text and visual features for the images in VerSe. Once the textual and visual features are projected, we merge them to get the multimodal representation. We experimented with two ways of combining visual and textual features projected via CCA or DCCA, namely interpolation (see Equations (2.3) and (2.5)) and concatenation.

To evaluate our proposed method, we compare against the first sense heuristic (FS), which defaults to the sense listed first in the dictionary (where senses are typically ordered by frequency). This is a strong baseline which is known to outperform more complex models in traditional text-based WSD. In VerSe we observe skew in the distribution of the senses and the first sense heuristic is as strong as it is on text. We further report the performance of the most frequent sense heuristic (MFS), which assigns the most frequently annotated sense for a given verb in VerSe. Note that MFS is supervised (as it requires sense annotated data to obtain the frequencies), so it should be regarded as an upper limit on the performance of the unsupervised methods we propose (as is also the case in unsupervised WSD for text (Navigli, 2009)).

In the GOLD setting we find that for both types of verbs, textual representations based on image descriptions (C) outperform visual representations (CNN features). The text-based results compare favorably to the original Lesk algorithm (as described in Equation (2.1)), which performs at 30.7 for motion verbs and 36.2 for non-motion verbs in the GOLD setting.

Using GOLD annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	1812	70.8	86.2	54.6	73.3	**75.6**	58.3	66.6	74.7	73.8	50.5	**75.4**	74.0	52.4	66.3	68.3
Non-Motion	1698	80.6	90.7	57.0	**72.7**	**72.6**	56.1	66.0	72.2	71.3	53.6	71.6	70.2	57.3	59.8	55.1

Using PRED annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	1812	70.8	86.2	65.1	54.9	61.6	58.3	**72.6**	63.6	66.5	54.0	56.6	56.2	57.1	56.5	56.2
Non-Motion	1698	80.6	90.7	59.0	64.3	64.0	56.1	63.8	**66.3**	**66.1**	50.7	55.3	54.8	49.5	50.0	50.0

Table 2.4: Sense disambiguation scores for **gold-standard verbs**: accuracy scores for motion and non-motion verbs using different types of sense and image representations (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic drawn from training set). Model configurations that performed the best are shown in **bold**.

This improvement is clearly due to the use of word2vec embeddings. We also experimented with Glove vectors (Pennington et al., 2014) but observed that word2vec representations consistently achieved better results that Glove vectors. Note that CNN-based visual features alone perform better than gold-standard object labels alone in the case of motion verbs.

We also observed that adding visual features to textual features improves performance in some cases: multimodal features perform better than textual features alone both for object labels (CNN+O) and for image descriptions (CNN+C). However, adding CNN features to textual features based on both object labels and descriptions (CNN+O+C) results in a small decrease in performance. Furthermore, we note that CCA models outperform simple vector concatenation in case of GOLD setting for motion verbs, and overall DCCA performs considerably worse than concatenation. For CCA and DCCA we varied λ from 0.1 to 0.9 giving different importance to textual and visual features. The best performing scores achieved using weighted interpolation of textual and visual features with $\lambda = 0.5$

When comparing to our baseline and upper limit, we find that all GOLD models which use descriptions-based representations (except DCCA) outperform the first sense heuristic for motion-verbs (accuracy 70.8), but not for non-motion verbs (accuracy 80.6). As expected, both motion and non-motion verbs perform significantly below the most frequent sense heuristic (accuracy 86.2 and 90.7 respectively), which provides an upper limit for unsupervised approaches. Even in traditional sense disambiguation models most-frequent sense is a very strong baseline which often a lot of supervised models fail to beat. This is mainly observed due to reporting bias observed in the distribution of the senses.

We now turn to results obtained using object labels and image descriptions predicted by state-of-the-art automatic systems (PRED configuration). This is arguably a more realistic scenario, as it only requires images as input, rather than human-generated object labels and image descriptions (though object detection and image description systems are required instead). In the PRED setting, we find that textual features based on object labels (O) outperform both first sense heuristic and textual features based on image descriptions (C) in the case of motion verbs. Combining textual and visual features via concatenation improves performance for both motion and non-motion verbs. The overall best performance of 72.6 is obtained by combining CNN features and embeddings based on object labels and outperforms the first sense heuristic in case of motion verbs (accuracy 70.8).

Image	Descriptions	Objects
	A man holding a nintendo wii game controller. A man and a woman playing a video game. A man and a woman are playing a video game.	person, bassoon, violin fiddle, oboe, hautboy
play: perform or transmit music, engage in competition		
	A woman standing next to a fire hydrant. A woman walking down a street holding an umbrella. A woman standing on a sidewalk holding an umbrella.	person, horizontal bar, high bar, pole
swing: move in a curve or arc, hang freely		
	A couple of cows standing next to each other. A cow that is standing in the dirt. A close up of a horse in a stable	arabian camel, dromedary, person
feed: give food, eat, be sustained on		

Table 2.5: Images assigned an incorrect sense (shown in red) in the PRED setting. In many of these cases predicted descriptions were not relevant to the image. Gold-standard senses are shown in blue.

Figure 2.9: We present the example grouping of verb *play* senses based on the predicted sense labels.First group of 2 images is labeled with *playing sport* sense, second group *playing instrument* and third group is labeled with *children playing* sense

In the PRED setting for both classes of verbs the simpler concatenation model performs better than the more complex CCA and DCCA models. Note that for CCA and DCCA we report the best performing scores achieved using weighted interpolation of textual and visual features with $\lambda = 0.3$. Overall, our findings are consistent with the intuition that motion verbs are easier to disambiguate than non-motion verbs, as they are more depictable and likely to involve objects. This is also reflected in the higher inter-annotator agreement for motion verbs (see Table 2.2).

In order to better understand where the proposed unsupervised algorithm fails, we analyzed images that were disambiguated incorrectly. In the PRED setting, we observed that automatically generated image descriptions obtained lower scores compared to predicted object labels. The main reason for this is that the generated descriptions are often unrelated to the action depicted, whereas the object labels predicted by the CNN model are mostly topical and related to the image. This highlights that current image description systems still have clear limitations, despite high evaluation scores reported in the literature (Vinyals et al., 2015; Fang et al., 2015b). Examples of images which were assigned incorrect senses are shown in Table 4.7 together with automatically generated descriptions and object labels. We also present images grouped according to

Motion verbs: 19, MFS: 76.1				
Features	GOLD		PRED	
	Sup	Unsup	Sup	Unsup
FS	60.0	60.0	60.0	**60.0**
O	82.3	35.3	80.0	43.8
C	78.4	53.8	69.2	41.5
O+C	80.0	55.3	70.7	45.3
CNN	82.3	58.4	82.3	58.4
CNN+O	**83.0**	48.4	**83.0**	**60.0**
CNN+C	82.3	66.9	82.3	53.0
CNN+O+C	**83.0**	58.4	**83.0**	55.3

Non-Motion Verbs: 19, MFS: 80.0				
Features	GOLD		PRED	
	Sup	Unsup	Sup	Unsup
FS	71.3	**71.3**	71.3	**71.3**
O	79.1	48.6	78.2	46.0
C	79.1	53.9	77.3	61.7
O+C	79.1	66.0	77.3	55.6
CNN	**80.0**	55.6	**80.0**	55.6
CNN+O	**80.0**	56.5	**80.0**	52.1
CNN+C	**80.0**	56.5	**80.3**	60.0
CNN+O+C	**80.0**	59.1	**80.0**	55.6

Table 2.6: Accuracy scores for motion and non-motion verbs for supervised and unsupervised approaches using different types of sense and image representation features (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Configurations that perform the best are shown in **bold**

same sense labels in Figure 2.9.

2.6.2 Supervised Experiments

We also investigated disambiguation performance in a supervised setting. Specifically, we trained logistic regression classifiers for sense prediction by dividing the images in VerSe into training and testing. To train the classifiers (one per verb), we selected verbs which have at least 20 images and at least two senses in VerSe. Few verbs such as *board*, *hang* only had one sense annotated in VerSe. Few other verbs have very skewed distribution of senses resulting in 5 or less number of images per sense. We ignore all such verbs. This resulted in 19 motion verbs and 19 non-motion verbs. The classifiers used textual (O, C) and visual (CNN) features, either in isolation or combined. Our results are summarized in Table 2.6; for comparison, we also report the scores of our unsupervised algorithm on the same set of verbs (in both GOLD and PRED settings).

We observe that supervised classifiers perform better than the first sense baseline (for both motion and non-motion verbs). In most cases multimodal features (CNN+C+O) outperform textual or visual features alone especially in the PRED setting, which is arguably the more realistic scenario. The features from PRED image descriptions show

Verb type	Verbs	Images	Accuracy		mAP	
			MLC	MIL	MLC	MIL
Motion	39	1,812	46.96	**50.60**	35.81	**41.47**
Non-motion	51	1,698	34.82	**37.47**	31.12	**35.27**

Table 2.7: Verb prediction accuracy and mAP on VerSe; MIL: Multiple Instance Learning; MLC: Multi-label classification.

better results for non-motion verbs for both supervised and unsupervised approaches, whereas PRED object features show better results for motion verbs. We also find that supervised classifiers outperform the most frequent sense for motion verbs, whereas for non-motion verbs our scores match the most frequent sense heuristic. When analysed the performance by verb we observe motion verbs such as *ride*, *run*, *serve* and *catch* are the best performing with supervised classifiers with greater than 90% accuracy. Among the non-motion verbs *point*, *stick* and *learn* performed the best.

2.6.3 Verb Prediction and Sense Disambiguation

We measure verb prediction performance using both accuracy and mean average precision (mAP). If a verb is used in at least one of the gold-standard image descriptions, it is included as a positive instance; as a result, an image can have multiple gold-standard verb labels. Both MLC and MIL systems output a distribution of verbs given an image. We consider verbs with probability higher than a threshold $\tau = 0.2$ as positive predictions.

Table 2.7 summarises the performance of MLC and MIL. As can be seen, MIL performs best both in terms of accuracy and mAP, across motion and non-motion verbs. Among motion verbs, the most accurately predicted ones were *drive, fly, ride, play*; for non-motion verbs *sit* and *hold* were most accurate. In Figure 2.10 we show top 3 verbs predicted by the MIL and MLC models for three different images. In Figure 2.11 we also show the visualizations of different senses of the verb play, which indicate that depending on the sense of verb being depicted our models are localizing different aspects of the image. In Figure 2.12 we present visualisations of different images for the verb *ride*, despite having different type of objects that people ride (animals, vehicles, boat etc.), our models predict the verb ride and localise the most relevant region of the image. Visualisations of verbs *fly*, *smile* and *feed* are presented in Appendix A.3.

Finally, Table 2.8 provides examples of the best and worst performing verbs for MLC and MIL using average precision (AP). Although informative, AP is a pessimistic evaluation metric because we can not exhaustively annotate all possible verbs depicted in an image. Consider the case where our model predicts the verbs *stand*, *hold*, *[play]* for an image depicting a person playing tennis. The predictions are all correct, but AP would penalize us if those verbs are not in our gold-standard annotation.

play, hit, serve

hold, play, hit

play, perform, sing

play, perform, sing

sit, play, hold

sit, hold, play

Figure 2.10: Example verb predictions of MIL and MLC classifiers

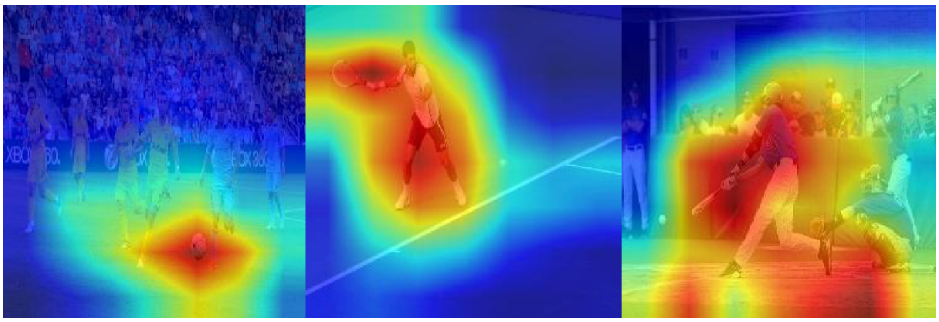
Verb	Count	MLC	MIL	Verb	Count	MLC	MIL
shoot	339	0.14	**0.16**	blow	1329	31.88	**44.05**
drill	128	0.26	**0.27**	draw	985	50.37	**63.27**
break	794	**2.26**	1.63	hit	6459	**68.98**	68.53
lift	980	3.89	**3.98**	kick	1780	75.00	**79.27**
chase	745	4.35	**5.05**	paddle	1027	76.41	**83.76**
stick	2948	7.14	**8.09**	fly	13395	80.90	**85.19**

Table 2.8: Average precision scores for individual verbs. Count refers to number of positive training instances. Verbs with the lowest and highest performance are shown.

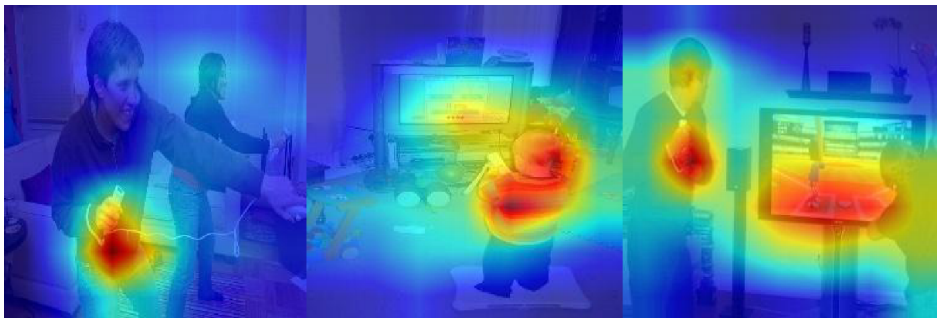
	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.15$	$\tau = 0.2$	$\tau = 0.25$	$\tau = 0.3$
Majority	48.5	57.6	63.5	66.6	**66.9**	64.6
All	68.2	74.8	78.5	**80.6**	80.3	76.0

Table 2.9: Human evaluation accuracy scores for verb prediction labels. τ is the confidence threshold of verb predictions.

(a) play instrument



(b) play sport



(c) children play video games

Figure 2.11: Localizations for different senses of the verb *play*.

2.6.4 Human Evaluation of Verb Prediction

To study in more detail the quality of the verb predictions, we conducted a human evaluation study. We presented the top 10 verbs predicted by the MIL classifier for a given image to Amazon Mechanical Turk workers and asked them to select those that apply. For this study, we sampled 640 images from VerSe across verbs and senses with 2–5 images per unique verb sense. For every image, we collected annotations from three workers. Overall, 54 workers took part in the study, with pair-wise inter-annotator agreement (ITA) of 0.741.

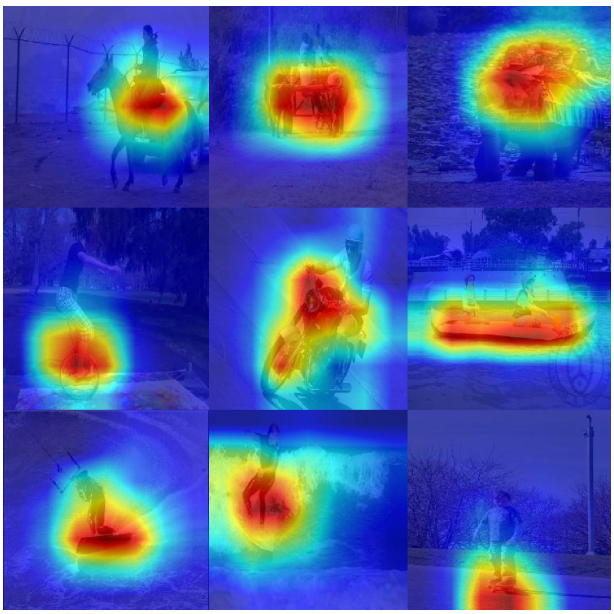


Figure 2.12: Localizations for the verb *ride* from different images highlighting the regions that contribute to the prediction of the verb.

In Table 2.9, we present mean accuracy scores across all 640 images using human selected verbs as gold-standard labels. Specifically, we compute accuracy for every image based on (a) majority labels, i.e., if at least two out of three annotators agreed that a particular verb is depicted in the image; and (b) all labels, i.e., if at least one annotator thought a particular verb is depicted in the image. The average number of verbs selected per image is 4.17 for majority labels and 6.18 for all labels. In Table 2.9 we present the accuracy scores against the gold-standard from the human annotation whilst we vary τ, the prediction confidence threshold. As can be seen, the best accuracies are achieved at $\tau = 0.2$ and $\tau = 0.25$. Overall, most verb predictions are considered appropriate by humans, even under the stricter majority label criterion.

Sense accuracy scores for predicted verbs are shown in Table 2.10. Again, scores are shown for motion and non-motion verbs separately. We report results for unsupervised methods, using the multiple instance learning approach to obtain verb predictions. Here, we only consider images for which the MIL system predicted the same verbs as in VerSe. These are 918 images compared to 1,812 in the full dataset. For this reason, we do not report supervised experiments in the predicted verb setting: there are not enough image-verb instances to train a supervised classifier.

Using GOLD annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	918	68.4	87.3	58.3	78.7	**82.7**	65.1	73.5	79.4	79.6	54.0	75.9	75.8	56.4	72.0	75.9
Non-Motion	637	83.8	92.3	63.7	78.1	**80.5**	58.7	73.3	76.9	76.7	59.6	73.4	70.1	61.9	63.1	61.2

Using PRED annotations for objects and captions

	Images	FS	MFS	Textual			Vis	Concat (CNN+)			CCA (CNN+)			DCCA (CNN+)		
				O	C	O+C	CNN	O	C	O+C	O	C	O+C	O	C	O+C
Motion	918	68.4	87.3	72.3	65.1	71.6	65.1	**79.4**	74.0	75.8	49.3	60.3	57.8	64.0	66.4	64.8
Non-Motion	637	83.8	92.3	65.7	**77.3**	76.2	58.7	70.0	74.4	74.2	49.6	59.1	59.1	54.0	53.0	54.6

Table 2.10: Sense disambiguation scores for **predicted verbs**: accuracy scores for motion and non-motion verbs using different types of sense and image representations (O: object labels, C: image descriptions, CNN: image features, FS: first sense heuristic, MFS: most frequent sense heuristic). Model configurations that performed the best are shown in **bold**.

Also notice that even though several of the verbs predicted by the MIL system may be appropriate for VerSe images, we do not have sense annotations for them to perform either evaluation or training. Overall, sense disambiguation results for predicted verbs follow the same pattern as those obtained from observed verbs: motion verbs are easier to disambiguate than non-motion ones; in the GOLD setting best model performance is achieved with object labels and image descriptions combined, whereas in the PRED setting concatenation of CNN features with object labels yields best results.

2.7 Conclusions

In this chapter, we introduced the new task of visual verb sense disambiguation: given an image and a verb, identify the verb sense depicted in the image. We developed VerSe, a new dataset with verb sense annotation based on the COCO and TUHOI datasets. We evaluated supervised and unsupervised visual sense disambiguation models and demonstrated that both textual and visual information associated with an image can contribute to sense disambiguation. In an in-depth analysis of various image representations we showed that object labels and visual features extracted using state-of-the-art convolutional neural networks result in good disambiguation performance, while automatically generated image descriptions were shown to be less useful.

We also explored a second scenario for visual sense disambiguation, where we assumed that only the image is given, and both the verb and its sense need to be predicted. We conceptualized this as a two-stage process: First, we predicted verb labels using multi-instance learning or multilabel classification. Then, we disambiguated the predicted verbs using our sense disambiguation approach combining visual and textual features. We showed that the verbs predicted by this method agree well with human intuitions, and we also obtained good sense accuracy scores. Note that the second scenario differs from our first scenario in a crucial respect: we are able to predict multiple verbs per image, and each of these verbs can be associated with a different image region (if the multi-instance learning model is used). While a lot of images in our dataset only depict a single action, this is not always the case (e.g., the child in the rightmost image in Figure 2.10 is both sitting in the sand and holding a toy).

In this chapter, we explored visual sense disambiguation in a language. An important area for future research is the connection between verb sense ambiguity and translation ambiguity. This rests on the observation that sense ambiguity in one language can manifest itself as ambiguity in lexical choice in another language. The English verb

ride, for instance, can have the senses (1) sit on and control a vehicle (as in *ride bicycle*), or (2) sit and travel on the back of animal (as in *ride hose*). These two senses corresponds to two different lexical choices in German, viz., the verbs *fahren* (for sense 1) and *reiten* (for sense 2). In other words, we need to sense disambiguate the verb in order to translate it correctly. This observation is of practical importance, as machine translation systems often suffer from disambiguation errors such as this (Vilar et al., 2006).

If the ambiguous verb occurs in a visual context, then we can apply the VSD methods developed in this chapter to the resolution of translation ambiguities as they occur in multilingual image description or crosslingual image retrieval. In Chapter 4 we discuss and demonstrate the applicability of visual sense disambiguation models to down streams tasks such as multimodal machine translation.

Chapter 3

Verb Prediction Models against Human Eye-tracking Data

In the previous chapter, we have developed models for predicting and disambiguating verbs from images. Our analysis has shown that these models focus on a particular region of the image in order to predict the verb. This is intuitively similar to how humans do the task of recognizing actions in the images. In this Chapter, we ask whether the predictions made by such models correspond to human intuitions about visual verbs or actions. We show that the image regions a verb prediction model identifies as salient for a given verb correlate with the regions fixated by human observers performing a verb classification task. Additionally, we also compare the correlation of human fixations against visual saliency models, center bias and model combinations.

3.1 Motivation

Humans can easily process visual data such as images and videos and extract informative regions suitable for high level cognitive tasks. There has been a large body of research in cognitive science on the connection between human eye movements and the way humans interpret and categorize images (Jaimes et al., 2001; Henderson, 2003). Therefore, there has been significant amount of research on using human gaze information for many computer vision tasks such as object detection (Yun et al., 2013; Papadopoulos et al., 2014), face and text detection (Karthikeyan et al., 2013), action recognition in images and videos (Vig et al., 2012; Ge et al., 2015; Dorr & Vig, 2017) and image segmentation (Mishra et al., 2009). Few recent studies have also combined low-level image features with human gaze information for action recognition (Ge et al., 2015) and using gaze information for identifying salient regions in videos for which visual features are then computed for action recognition and localization.

In this chapter, we take this a step further and analyse whether a verb prediction model, which labels an image with potential verbs, correlates with human fixations. Our automatic and human evaluation experiments show that convolutional neural network based verb prediction models achieves good verb prediction accuracy. However, it is not clear to what extent the model captures human intuitions about visual verbs. Specifically, it is interesting to ask whether the image regions that the model identifies as salient for a given verb correspond to the regions a human observer relies on when determining which verb is depicted. The output of a verb prediction model can be visualized as a heatmap over the image, where hot colors indicate the most salient areas for a given task (see Figure 3.5 for examples). In the same way, we can determine which regions a human observes attends to by eye-tracking them while viewing the image. Eye-tracking data consists of a stream of gaze coordinates, which can also be turned into a heatmap. Model predictions correspond to human intuitions if the two heatmaps correlate.

We show that the heatmaps generated by the verb prediction model in Section 2.5 correlate well with heatmaps obtained from human observers performing a verb classification task. We achieve a higher correlation than a range of baselines: center bias, visual salience, and model combinations, indicating that the verb prediction model successfully identifies those image regions that are indicative of the verb depicted in the image.

3.2 Related Work

Human eye movement data which provides insights into where humans fixate their eyes when performing a task has been shown to be useful in various computer vision tasks (Vig et al., 2012; Winkler & Subramanian, 2013; Karthikeyan et al., 2013; Papadopoulos et al., 2014). Human eye movement data has been used to improve standard face and text detection algorithms by identifying the target regions of interest from human fixations (Karthikeyan et al., 2013). Similarly, fixation data collected while humans perform the task of finding a target object in an image has shown to be useful for detecting object bounding boxes in images (Papadopoulos et al., 2014). Using eye tracking to train object detectors has been argued to reduce annotation time since eye tracking data could be collected from naive viewers where as bounding box annotations require annotation guidelines and expert annotators.

Eye movement data has been extensively used to track and identify salient regions and enhance action classifiers for videos and images (Mathe & Sminchisescu, 2012; Vig et al., 2012; Mathe & Sminchisescu, 2013). Most of the studies on human eye movement data for action classification is targeted for videos. Winkler & Subramanian (2013) present an overview of existing human eye tracking datasets. The PASCAL VOC Actions Fixations Dataset is only the large scale human fixation dataset available for still images (Mathe & Sminchisescu, 2013). This dataset has been used to understand and predict gaze patterns and visual scanpaths for action classification. Recent studies include complex methods of combining eye tracking fixations with state-of-the-art classification convolutional neural network algorithms to improve action classification performance (Ge et al., 2015).

While most of the prior work has used human eye movement data to build or augment existing models for various tasks, very few of them have actually compared model localizations with human fixations. The most closely related analysis to ours is the work by Das et al. (2016) who tested the hypothesis that the regions attended to by neural visual question answering (VQA) models correlate with the regions attended to by humans performing the same task. Their results were negative: the neural VQA models do not predict human attention better than a baseline visual salience model (see Section 3.4). It is possible that this result is due to limitations of the study of Das et al. (2016): their evaluation dataset, the VQA-HAT corpus, was collected using mouse-tracking, which is less natural and less sensitive than eye-tracking. Also, their participants did not actually perform question answering, but were given a question and

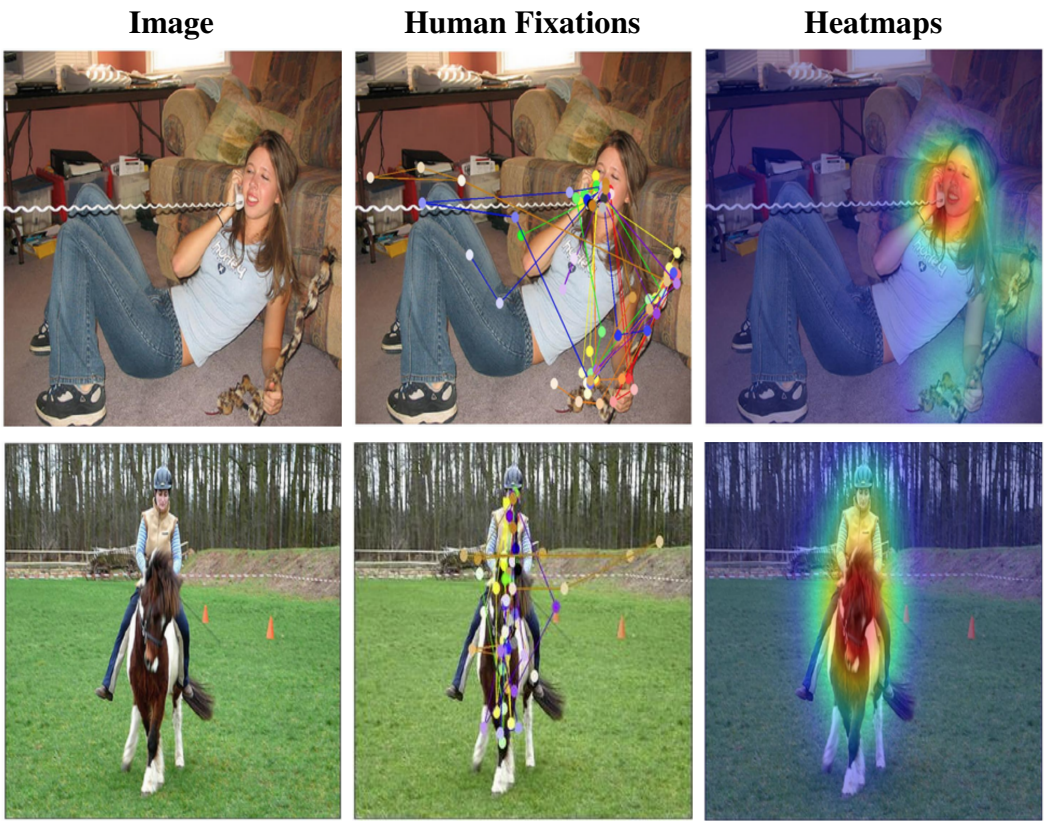


Figure 3.1: Example images with human fixations and heatmap visualizations of human fixations. Different colors in human fixations indicate fixations collected from different observers.

its answer, and then had to mark up the relevant image regions. Das et al. (2016) report a human-human correlation of 0.623, which suggests low task validity.

Qiao et al. (2018) also use VQA-HAT, but in a supervised fashion: they train the attention component of their VQA model on human attention data. Not surprisingly, this results in a higher correlation with human heatmaps than Das et al. (2016) unsupervised approach. However, Qiao et al. (2018) fail to compare to a visual salience model (given their supervised setup, such the salience model would also have to be trained on VQA-HAT for a fair comparison).

The work that is perhaps closest to our own work is Hahn & Keller (2016), who use a reinforcement learning model to predict eye-tracking data for text reading (rather than visual processing). Their model is unsupervised (there is no use of eye-tracking data at training time), but achieves a good correlation with eye-tracking data at test time. Furthermore, a number of authors have used eye-tracking data collected for text reading to train models that perform part-of-speech tagging (Barrett et al., 2016a,b,

2018), grammatical function classification (Barrett & Søgaard, 2015), and sentence compression (Klerke et al., 2016).

3.3 Eye-tracking Dataset

Human eye movement data have been studied for various tasks involving both images and videos. Most of the existing datasets for action recognition were targeted at videos (Rodriguez et al., 2008; Marszalek et al., 2009) and have been collected for small scale studies (See (Winkler & Subramanian, 2013) for overview). We use the PASCAL VOC Actions Fixation dataset as it is the largest task-controlled eye tracking dataset available for still images.

Mathe & Sminchisescu (2013) created the PASCAL VOC Actions Fixation dataset by annotating each image in the PASCAL VOC visual recognition challenge dataset with the eye-fixations of eight human observers who, for each image, were asked to recognize the action depicted and respond with one of the PASCAL VOC action class labels. The PASCAL VOC is a visual recognition challenge widely known in the computer vision community for evaluating performance on tasks such as object category detection and action classification. The PASCAL VOC Actions dataset (Everingham et al., 2010) contains 9,157 images covering 10 action classes (phoning, reading, jumping, running, walking, riding bike, riding horse, playing instrument, taking photo, using computer).

All the eye movements in the dataset were recorded using an SMI iView X HiSpeed 1250 tower-mounted eye tracker, with a sampling frequency of 500Hz. The LCD display had a resolution 1280×1024 pixels, with a physical screen size of 47.5×29.5 cm. Before displaying each image, all the subjects were required to fixate a target in the center of a uniform background on the screen. All the subjects were asked to perform a multi target detect and classify task i.e., press a key each time they have identified a person performing an action from the given set of PASCAL VOC action classes. Multiple choice options selected were recorded through a set of check-boxes displayed after each subject was given a three seconds to freely view an image while the x- and y-coordinates of their gaze positions were recorded. (Note that the original dataset also contained a control condition in which four participants performed visual search; we do not use the data from this control condition.) In Figure 3.1 we show examples of images from the PASCAL VOC Actions Fixation dataset with aggregated fixations from different subjects and heatmaps weighted by fixation duration. IThe fixation dataset

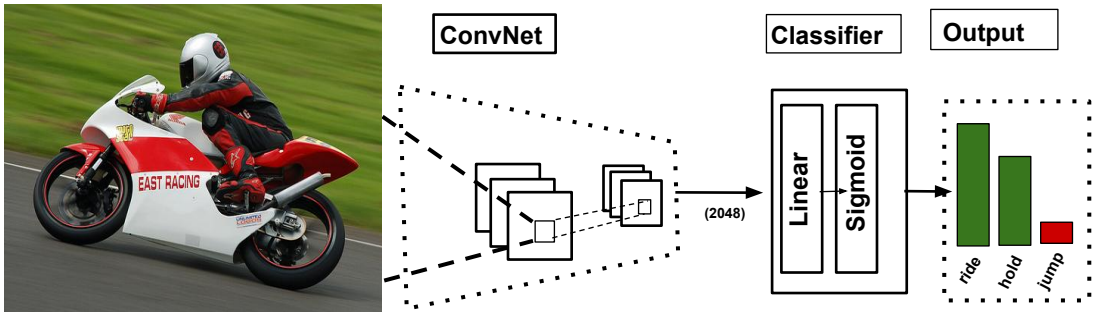


Figure 3.2: A schematic view of our multilabel verb classification model.

contains 1,085,381 fixation in total and the the average scanpath length for action subjects is 10.0 including the initial central fixation.

As discussed in Chapter 2 actions and verbs are distinct concepts (Ronchi & Perona, 2015; Pustejovsky et al., 2016; Gella & Keller, 2017), we can still use the PASCAL Actions Fixation data to evaluate our model. When predicting a verb, the model presumably has to attend to the same regions that humans fixate on when working out which action is depicted – all the actions in the dataset are verb-based, hence recognizing the verb is part of recognizing the action.

3.4 Fixation Prediction Models

3.4.1 Verb Prediction Model (M)

In our study, we used the multi-label verb prediction model proposed in Chapter 2, which employs a multilabel CNN-based classification approach and is designed to simultaneously predict all verbs associated with an image. This model is trained over a vocabulary that consists of the 250 most common verbs in the TUHOI, Flickr30k, and COCO image description datasets (Le et al., 2014; Young et al., 2014b; Lin et al., 2014). For each image in these datasets, we obtained a set of verb labels by extracting all the verbs from the ground truth descriptions of the image (each image comes with multiple descriptions, each of which can contribute one or more verbs).

This model uses a sigmoid cross-entropy loss and the ResNet 152-layer CNN architecture. The network weights were initialized with the publicly available CNN pretrained on ImageNet and finetuned on the verb labels. The model architecture is shown schematically in Figure 3.2.

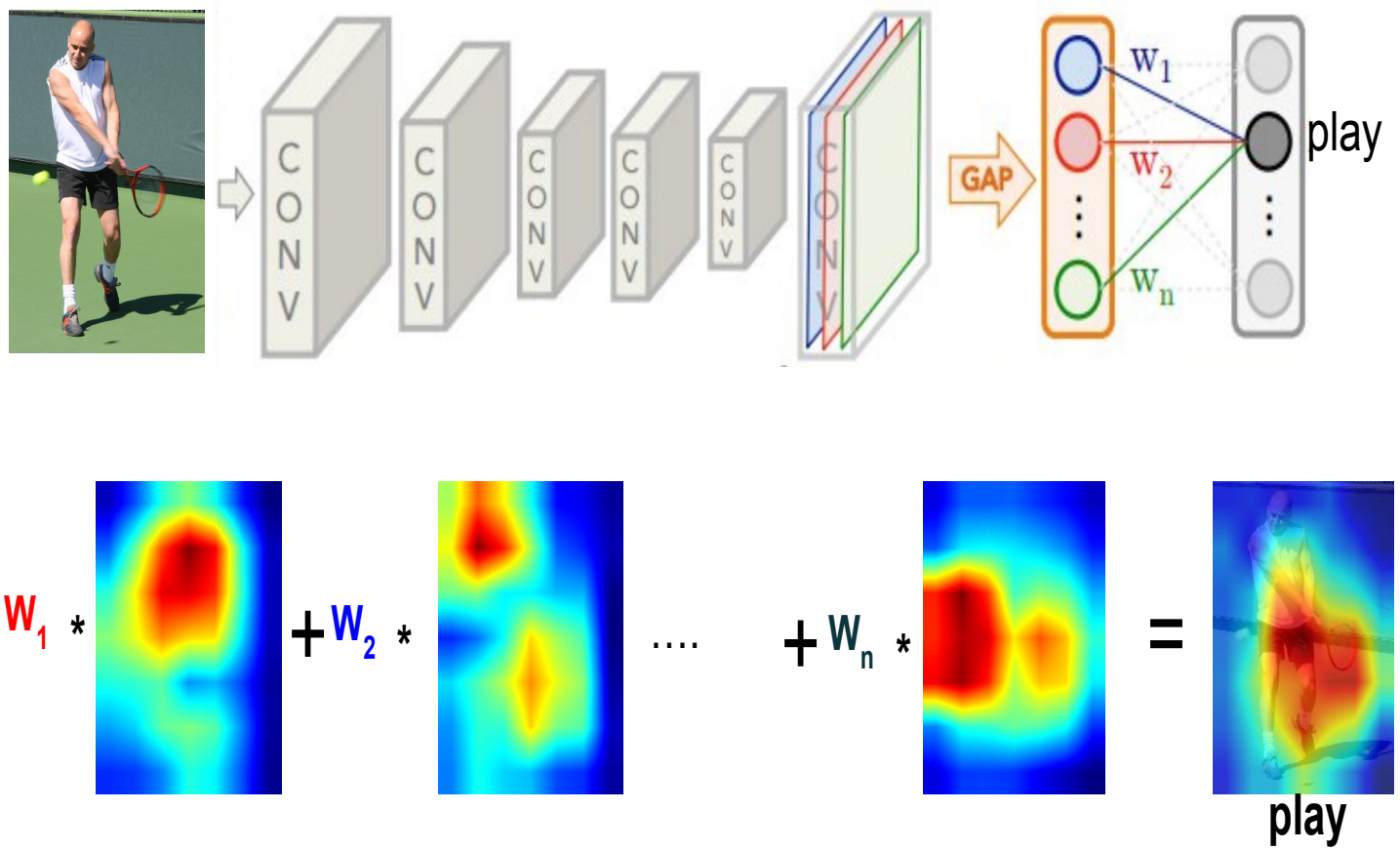


Figure 3.3: Visualisation of Class Activation Mapping

3.4.2 Class Activation Mapping (CAM)

To derive fixation predictions, we turned the output of the verb prediction model into heatmaps using the class activation mapping (CAM) technique proposed by Zhou et al. (2016). CAM uses global average pooling (GAP) of convolution feature maps to identify the important image regions by projecting back the weights of the output layer onto the convolutional feature maps using:

$$hm_c = \sum_k W_k f_k \tag{3.1}$$

where hm_c gives the salient features used in classifying the image as action class c, f_k denotes the kth feature map, W_k correspond to the weight of classification layer for feature map k leading to action class c. A Visualisation of class activation mapping is shown in Figure 3.3. This technique has been shown to achieve competitive results on both object localization and localizing the discriminative regions for action classification (Zhou et al., 2016).

3.4.3 Center Bias (CB)

We compare against a center bias baseline, which simulates the task-independent tendency of observers to make fixations towards the center of an image. In many free-viewing eye tracking tasks, human subjects choose to direct their initial fixations toward the center of the image (Renninger et al., 2007). Center bias is considered as a strong baseline for most eye-tracking datasets (Tatler, 2007). We follow Clarke & Tatler (2014) and compute a heatmap based on a zero mean Gaussian with a co-variance matrix:

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & v\sigma^2 \end{pmatrix}$$

where $\sigma^2 = 0.22$ and $v = 0.45$, where v captures the horizontal bias present in the human data. These σ and v values suggested by Clarke & Tatler (2014) based on an evaluation on 10 different eye tracking datasets. This Gaussian distribution is treated as an attention map and compared against the attention maps of human eye fixations.

3.4.4 Visual Salience (SM)

Models of visual salience are meant to capture the tendency of the human visual system to fixate the most prominent parts of a scene, often within a few hundred milliseconds

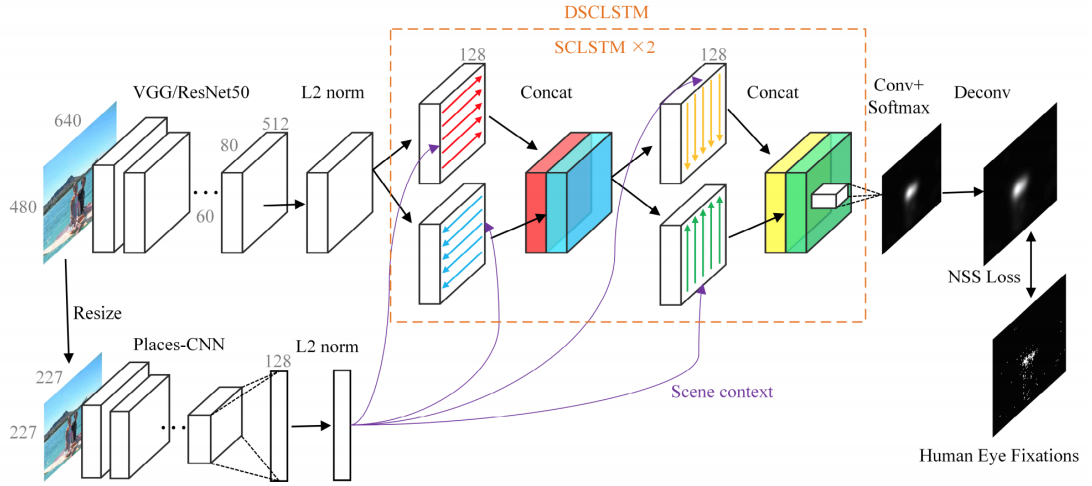


Figure 3.4: A deep spatial contextual long-term recurrent convolutional neural network (DSCLRCN) for saliency prediction

of exposure. A large number of salience models have been proposed in the literature to predict the salient regions of the image (Kümmerer et al., 2014; Liu & Han, 2018; Kruthiventi et al., 2016; Pan et al., 2016; Cornia et al., 2016). We chose the deep spatial contextual long-term recurrent convolutional network (DSCLRCN) of Liu & Han (2018) trained on SALICON (Jiang et al., 2015), a large human attention dataset, to infer salience for arbitrary images.

DSCLRCN model is designed to mimic the ability of the human visual processing system to incorporate the global context of the image to predict the saliency. DSCLRCN first learns powerful local saliency features on multiple regions distributed throughout the image. These local features features are extracted using pretrained CNN's on ImageNet and Places datasets (Deng et al., 2009; Zhou et al., 2014). Both ImageNet features and scene features are concatenated and fed to the DSCLSTM model which propagates global and contextual scene information to local image regions using a recurrent deep spatial long short-term network (shown in Figure 3.4). A final convolutional layer is used to obtain the saliency map with normalized saliency scanpath loss between the predicted saliency map and the ground truth human eye fixations.

Note that salience models are normally tested using free viewing tasks or visual search tasks, not verb prediction. However, salience can be expected to play a large role in determining fixation locations independent of task, so DSCLRCN is a good baseline to compare to. DSCLRCN model ranks among the top three models with one of the

highest correlation with human fixations on the MIT300 saliency benchmark out of 80 models (Bylinskii et al., 2016).

SALICON is a large dataset containing 20,000 images (a subset of the COCO dataset) each annotated with pixelwise semantic annotations of saliency annotations. Saliency annotations are collected using a popular substitute of eye-tracking annotations a crowd sourcing approach using mouse-tracking. A large body of saliency prediction research uses SALICON to train their models since it is the largest dataset available for saliency prediction.

3.5 Results

To evaluate the similarity between human fixations and model predictions, we first computed a heatmap based on the human fixations for each image. We used the PyGaze toolkit (Dalmaijer et al., 2014) to generate Gaussian heatmaps weighted by fixation durations. We then computed the heatmap predicted by our model for the top-ranked verb the model assigns to the image (out of its vocabulary of 250 verbs). We used the rank correlation between these two heatmaps as our evaluation measure. For this, both maps are converted into a 14×14 grid, and each grid square is ranked according to its average attention score. Spearman's ρ is then computed between these two sets of ranks.

For a pair of maps h_x and h_y, the rank correlation is then:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i is difference between the rank of values in the i^{th} square on the grid for h_x and h_y. This is the same evaluation protocol that Das et al. (2016) used to evaluate the heatmaps generated by two question answering models with unsupervised attention, viz., the Stacked Attention Network (Yang et al., 2016) and the Hierarchical Co-Attention Network (Lu et al., 2016b). This makes their rank correlations and ours directly comparable.

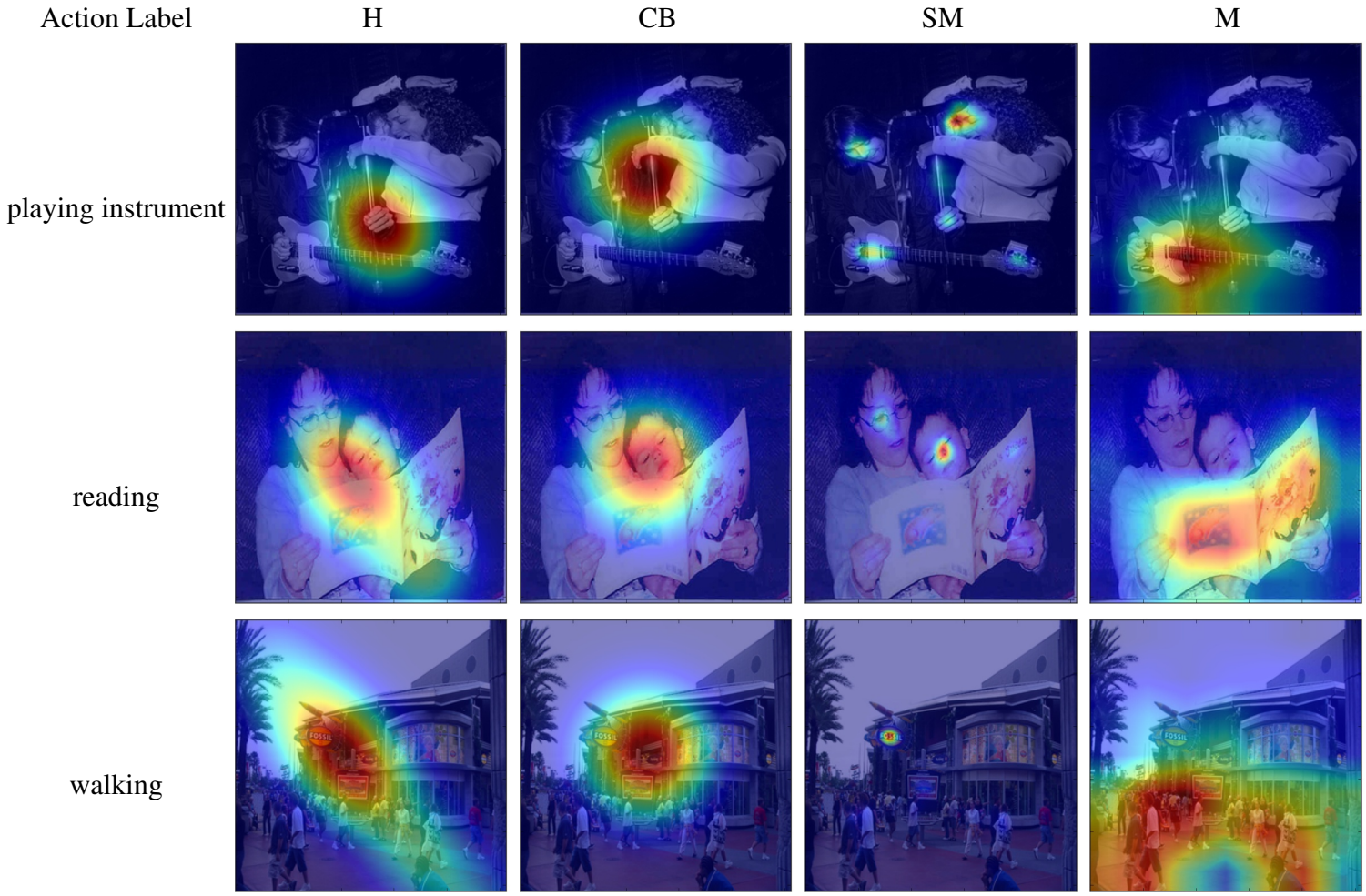


Figure 3.5: Heatmaps visualizing human fixations (H), Center Bias (CB), salience model (SM) predictions, and verb model (M) prediction for images depicting actions playing, reading and walking

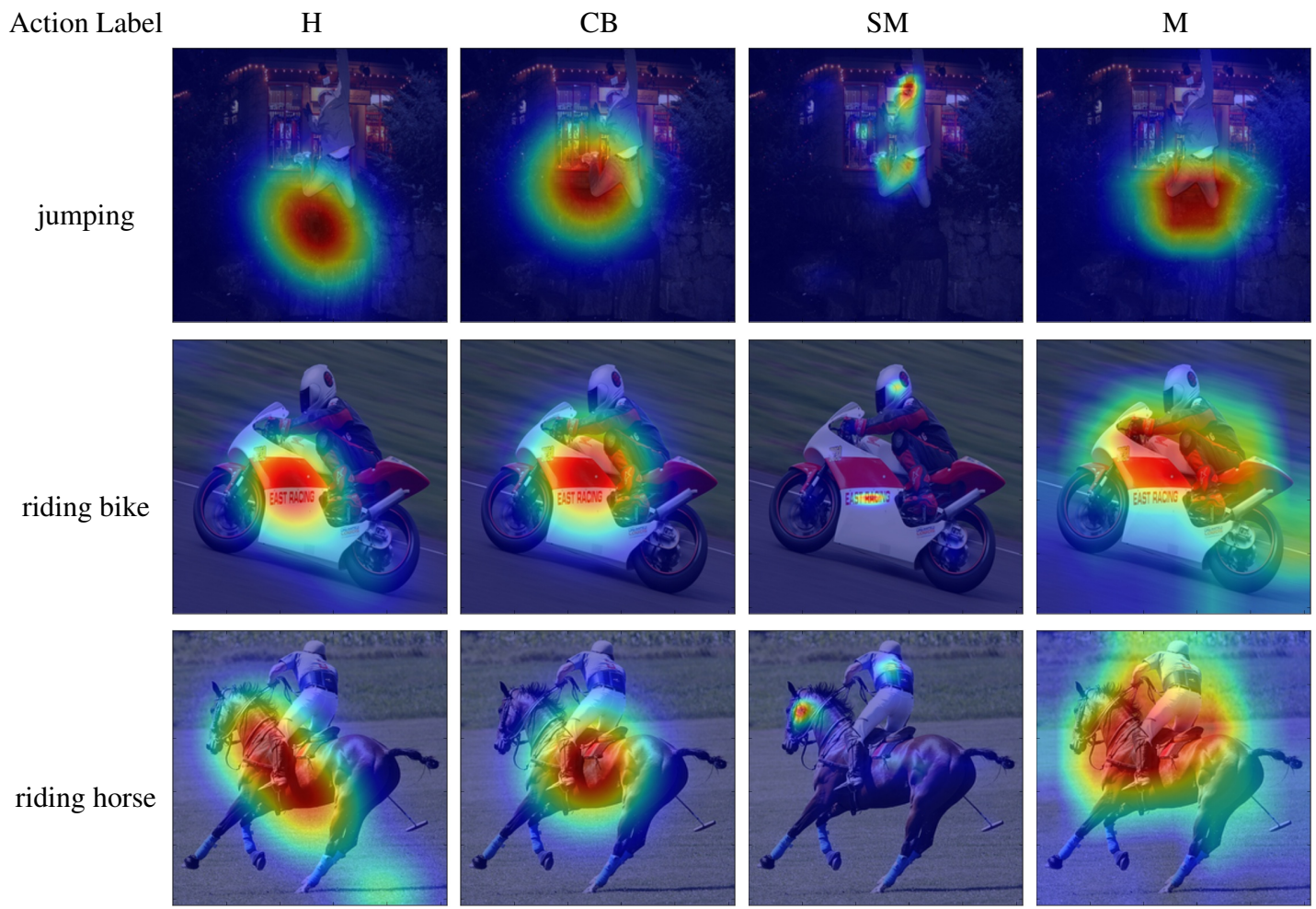


Figure 3.6: Heatmaps visualizing human fixations (H), Center Bias (CB), salience model (SM) predictions, and verb model (M) prediction for images depicting actions jumping, riding bike and riding horse

In Table 3.1 we present the correlations between human fixation heatmaps and model-predicted heatmaps. All results were computed on the validation portion of the PASCAL Actions Fixation dataset. We average the correlations for each action class (though the class labels were not used in our evaluation), and also present overall averages. In addition to our model results, we also give the correlations of human fixations with (a) the center bias baseline, and (b) the salience model. We also report the correlations obtained by all combinations of our model and these baselines. Finally, we report the human-human agreement averaged over the eight observes. This serves as an upper bound to model performance.

The results show a high human-human agreement for all verbs, with an average of 0.923. This is considerably higher than the human-human agreement of 0.623 that Das et al. (2016) report for their question answering ask, indicating that verb classification is a task that can be performed more reliably than Das et al. (2016) VQA region markup task (they also used mouse-tracking rather than eye-tracking, a less sensitive experimental method).

We also notice that the center baseline (CB) generally performs well, achieving an average correlation of 0.592. The salience model (SM) is less convincing, averaging a correlation of 0.344. This is likely due to the fact that SM was trained on the SALICON dataset; a higher correlation can probably be achieved by fine-tuning the salience model on the PASCAL Actions Fixation data. However, this would no longer be fair comparison with our verb prediction model, which was not trained on fixation data (it only uses image description datasets at training time, see Section 3.4). Adding SM to CB does not lead to an improvement over CB alone, with an average correlation of 0.591.

Our model (M) on its own achieves an average correlation of 0.529, rising to 0.628 when combined with center bias, clearly outperforming center bias alone (we use the mean pool of heatmap of center bias with our model predicted heatmap). Adding SM does not lead to a further improvement (0.626). The combination of our model with SM performs only slightly better than the model on its own.

In Figure 3.5 and Figure 3.6, we visualize samples of heatmaps generated from the human fixations, the center-bias, the salience model, and the predictions of our model. We observe that human fixations and center bias exhibit high overlap. Most of the images in PASCAL VOC dataset contain simpler scenes and there is high possibility of having the target object around the center of the image. The salience model attends to regions that attract human attention independent of task (e.g., faces), while our model mimics human observers in attending to regions that are associated with the verbs

		Rank correlations							
Verb	Images	H	CB	SM	M	CB+SM	CB+M	M+SM	M+CB+SM
phoning	221	0.911	0.599	0.361	0.562	0.598	**0.654**	0.569	0.652
reading	231	0.923	0.589	0.404	0.544	0.598	**0.655**	0.558	0.655
jumping	201	0.930	0.612	0.300	0.560	0.609	**0.650**	0.561	0.647
running	154	0.934	0.548	0.264	0.536	0.545	**0.604**	0.536	0.602
walking	195	0.938	0.553	0.311	0.535	0.552	**0.611**	0.537	0.609
riding bike	199	0.925	0.580	0.329	0.518	0.578	**0.622**	0.527	0.621
riding horse	206	0.910	0.593	0.351	0.532	0.588	**0.604**	0.532	0.601
playing instrument	229	0.925	0.571	0.350	0.478	0.568	**0.596**	0.484	0.593
taking photo	205	0.925	**0.656**	0.354	0.508	0.647	0.630	0.514	0.628
using computer	196	0.916	0.633	0.389	0.525	0.626	**0.655**	0.533	0.652
overall	2037	0.923	0.592	0.344	0.529	0.591	**0.628**	0.535	0.626

Table 3.1: Table of average rank correlation scores for the verb prediction model (M), compared with the upper bound of average human-human agreement (H), center bias (CB) baseline (Clarke & Tatler, 2014), and salience map (SM) baseline (Liu & Han, 2018). Results are reported on the validation set of the PASCAL VOC 2012 Actions Fixation data (Mathe & Sminchisescu, 2013). The best score for each class is shown in **bold** (except upper bound). Model combination are by mean of heatmaps.

depicted in the image. The saliency model heatmaps are very focused, which is a consequence of that model being trained on SALICON, which contains focused human attention maps.

For the *riding bike* example in Figure 3.6 humans fixate around the text on the bike which does not have any relation to the action riding bike. Our verb prediction model localizations are on the human and his position on the bike. We can also observe that our model predicted image regions vary with the different uses of a given verb (riding bike vs. riding horse). In the *playing instrument* example in Figure 3.6 humans fixated more around the flute whereas our model localizations are around the guitar (a more visible object).

3.6 Conclusions

We showed that a model that is trained to label images with verbs learned from image description data is able to predict which image regions humans attend when performing the task of assigning verb labels to images. The model therefore captures aspects of human intuitions about how verbs are depicted. This is an encouraging result given that our verb prediction model was not designed to model human behavior, and was trained on an unrelated MS COCO image description dataset, without any access to eye-tracking data (Lin et al., 2014). Our result contradicts the existing literature (Das et al., 2016), which found no above-baseline correlation between human attention and model attention in a VQA task. An explanation for this could be the use of less sensitive mouse-tracking data by as substitute for eye tracking data in evaluating VQA task. Another explanation could be visual question answering is more complicated task than simple action classification and CNN models are able to mimic human attention for simpler tasks.

Chapter 4

Cross-lingual Word Sense Disambiguation using Visual Context

Language is inherently ambiguous, and we often use perceptual information to resolve the ambiguity. Textual information available on the web is often found alongside images and we believe this visual information could be used to resolve ambiguity and improve various natural language processing applications. In the past few years, we have seen significant improvement in machine translation accuracy with neural machine translation models. However, the problem of lexical ambiguity in machine translation still remains unresolved especially for the part-of-speech categories such as verbs (Specia et al., 2016; Shah et al., 2016; Lala & Specia, 2018). To address lexical ambiguity in verbs across languages when visual information is available such as image tags or descriptions, we propose the task of cross-lingual visual sense disambiguation for verbs. Given a verb and an image as visual context, the task is to identify the correct translation of the verb in a target language. We develop the new MultiSense dataset, in which 9,504 images are annotated with one of 55 verbs and its translations in Spanish and German. We propose a series of cross-lingual visual sense disambiguation models and show that multimodal models that fuse textual information with visual features perform best on the task. We then demonstrate that visual sense disambiguation can be used to improve the performance of a standard unimodal machine translation system on image descriptions.

4.1 Motivation

Resolving lexical ambiguity remains one of the most challenging problems in natural language processing. It is often studied as a word sense disambiguation (WSD) problem: the task of assigning the correct sense to a word in a given context (Kilgarrif, 1998). Standard WSD disambiguates a word based on its *textual context* alone; however, in a multimodal setting *visual context* is also available and can be used for disambiguation.

Visual sense disambiguation for nouns (e.g., *mouse* can mean *small rodent* or *computer mouse*) can be performed with the help of an object detector trained on a large image database organized by word senses (Barnard & Johnson, 2005; Loeff et al., 2006; Saenko & Darrell, 2008), such as ImageNet. Visual sense disambiguation for verbs is more challenging (Gella et al., 2016), as no equivalent image database exists for verbs, and actions are harder to detect than objects, as they do not have clearly defined spatial boundaries.

Most WSD approaches obtain lists of possible word meanings from sense inventories such as WordNet (Miller, 1995). As an alternative, one can use the fact that a given word often has more than one translation into another language, and these translations can stand in for word senses (Carpuat & Wu, 2007; Navigli, 2009). As an example consider the verb *ride*, which can translate into German as *fahren* (ride a bike) or *reiten* (ride a horse). Prior work on cross-lingual WSD has been limited in scale and has only employed textual context (Lefever & Hoste, 2013).

The aim of this chapter is to bring multimodality and cross-linguality together: we propose to perform visual sense disambiguation using verb translations. The task is, given a verb and an image as visual context, to identify the correct translation of the verb in a target language. We present the new MultiSense dataset, in which 9,504 images are annotated with one of 55 verbs and its translations in Spanish and German. We propose a series of supervised cross-lingual visual sense disambiguation models. This includes unimodal models that use either textual or visual context, and multimodal models that combine the two via early or late fusion. We experiment with textual context generated by image description and situation recognition systems. Our results show that unimodal models yield good performance for the task, but are always outperformed by multimodal models. We also find that early fusion works better than late fusion, and that situation recognition generally provides better textual context than image description.

Source	Three guys **riding** on an elephant with a house-like structures and trees in the background.	A woman in a black dress and hat **rides** a unicycle in front of a crowd.
Ref	Drei Männer **reiten** auf einem Elefanten mit hausähnlichen Gebäuden und Bäumen im Hintergrund.	Eine Frau in einem schwarzen Kleid und mit Kopfbedeckung **fährt** vor Zuschauern auf einem Einrad.
NMT	Drei Jungs **fahren** auf einem Elefanten mit hausähnlichen Strukturen und Bäumen im Hintergrund.	Eine Frau in einem schwarzen Kleid und Hut **reitet** vor einer Menschenmenge.

Figure 4.1: Examples of errors made by the English-German NMT system of (Sennrich et al., 2017).

Cross-lingual visual sense disambiguation is an interesting task in its own right, but it also has a clear application in machine translation (MT). Getting the verb right is crucial for high quality MT output, and sometimes unimodal MT systems fail when the correct translation would be obvious from visual information (see Figure 4.1 for examples involving *ride*). Therefore, it should be possible to improve the performance of a unimodal MT system by using cross-lingual visual sense disambiguation to guide it to the correct verb. To test this claim, we annotate part of our MultiSense dataset with English image descriptions and their German translations. We then use the verbs returned by our visual sense disambiguation model to constrain the output of a neural MT system and demonstrate a clear improvement in Meteor, BLEU, and verb accuracy over a unimodal baseline.

4.2 Related Work

Multilingual tasks in which visual information is useful include bilingual lexicon learning and machine translation. Recent work has shown promising results for bilingual lexicon induction using images as a pivot or by combining visual information with

English	Translation1	Translation2
painting	malen	streichen
swinging	schaukeln	schwingen
brushing	putzen	bürsten
driving	fahren	treiben
drawing	zeichnen	gezogene

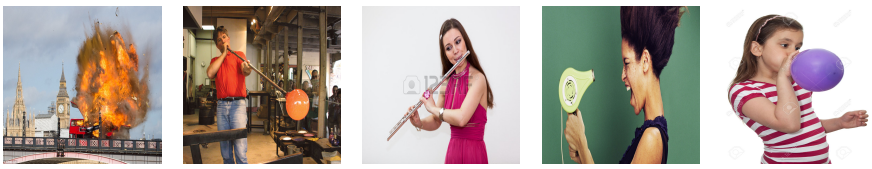
Table 4.1: Words with multiple translations from English to German that depend on context

cross-lingual vector spaces (Bergsma & Van Durme, 2011; Kiela et al., 2015; Vulic et al., 2016). However, as with other grounding or word similarity tasks, bilingual lexicon induction has so far mainly targeted nouns. Recent work by (Hartmann & Søgaard, 2018) show that existing methods that rely on images extracted from web for learning bilingual representations achieve lower scores when applied to other categories.

Similarly, most of the work on language grounding is either for image descriptions (Young et al., 2014b) or noun categories (Silberer & Lapata, 2014; Bruni et al., 2014; Kiela & Bottou, 2014; Lazaridou et al., 2014). An exception are action recognition or human object interaction detection tasks which are targeted at verbs. However, the labels used for these tasks often ignore basic verb semantic distinctions (Ronchi & Perona, 2015; Gella & Keller, 2017). For example, the action label "playing" cuts across the senses of the verb *play*, which include *play instrument*, *play sport*, and *engage in playful activity*.

In the context of multimodal machine translation, Hitschler et al. (2016) and Gella et al. (2017b) showed that visual context can be helpful for representation learning by using images as pivot between languages. However, their analysis reveals that most of the benefit derived from visual context is due to translating or disambiguating nouns. Other models proposed for multimodal machine translation have also shown better performance with visual information (Shah et al., 2016; Huang et al., 2016; Calixto et al., 2017a; Elliott & Kádár, 2017; Caglayan et al., 2018; Helcl et al., 2018), but did not investigate if these improvements hold for verbs. The present chapter addresses this issue by showing that visual WSD targeted specifically at verbs can boost MT performance over a unimodal baseline.

There are no existing datasets which specifically multimodal target translation ambiguity for verbs. For example, the SemEval cross-lingual WSD datasets (Lefever &



German	detonieren	blasen	blasen	föhen	aufblasen
Spanish	explotar	soplar	tocar	mandar	hinchar

Table 4.2: Example images from MultiSense for the verb *blow* annotated with verb translations in German and Spanish. Images correspond to the uses of *blowing up a bomb*, *blowing glass*, *blowing the flute*, *blowing with a hair dryer* and *blowing a balloon* respectively.

German	bürsten	putzan	fegen	striegeln	abbürsten
Spanish	peinar	cellipar	cellipar	cellipar	pintar

Table 4.3: Example images from MultiSense for the verb *brush* annotated with verb translations in German and Spanish. Images correspond to the uses of *brushing hair*, *brushing teeth*, *brushing floor*, *brushing horse* and *brushing bread* respectively.

Hoste, 2010) are unimodal, i.e., there is no visual context, and they only contain nouns. The VerSe dataset (Gella et al., 2016) includes verb/image pairs annotated with the word senses, but is monolingual, i.e., only covers English OntoNotes senses. The same holds for multimodal datasets annotated with semantic frames (Chao et al., 2015a; Yatskar et al., 2016). Multimodal translation datasets include Multi30k (Elliott et al., 2016, 2017), which provides German and French translations for 31k English image descriptions, but is not designed for WSD, and therefore includes only few ambiguous verbs. The Ambiguous COCO dataset released as part of the MMT 2017 task is a subset of our VerSe introduced in Chapter 2 (Gella et al., 2016), created in such a way to capture verb ambiguity in English. However, ambiguity in English could translate to ambiguity across languages in few cases. The recent Multimodal Lexical Translation (MLT) dataset by Lala & Specia (2018) is designed to include ambiguous words of all syntactic categories, but in practice 90% of them are nouns. Therefore, ours is the first one multimodal dataset that specifically targets translation-ambiguous verbs.

4.3 MultiSense Annotation

Our new MultiSense dataset pairs sense-ambiguous English verbs with images as visual context and contextually appropriate German and Spanish translations. To compile the dataset, we selected English verbs which had multiple translations into German and Spanish in Wiktionary[1], an online dictionary. A prior analysis of both Wiktionary and GermaNet (German WordNet) by Meyer & Gurevych (2010) showed Wiktionary has better coverage of translation mappings.

Verb Selection As our aim is to provide visual context for each verb, only visually depictable or concrete verbs can be used. A recent study by Hewitt et al. (2018) on large scale word translation via images showed gradual degradation in performance as words become more abstract. Most existing datasets such as VerSe (Gella et al., 2016), COCO-a (Ronchi & Perona, 2015), and imSitu (Yatskar et al., 2016) are built using depictable verbs. Among the verbs in these three datasets, we used those that have multiple Wiktionary translations for both German and Spanish, which resulted in a set of 122 English verbs. In this work, we only aim to address ambiguity across languages for visual verbs. However, this do not address all senses of verbs as in realisti scenario texts presented for Machine Translation could have non-visual senses as well.

A verb that is depictable in English can have multiple translations, but some of these may be non-depictable. For example, one of the translations of *change* in German is *ändern* (to become something different). We therefore asked native speakers of German and Spanish to annotate the Wiktionary translations as depictable or non-depictable (a similar method was used by Chao et al. (2015c) and Gella et al. (2016)). This resulted in a set of 55 English verbs which had multiple depictable translations in both German and Spanish. For these 55 verbs, the average number of Wiktionary translations in German and Spanish was 5.40 and 3.97, while the number of average visual translations in MultiSense was 3.18 and 3.01, respectively.

Images Paired with Verb Translations We retrieved candidate images by searching the web for verb phrases that included the target verb. These verb phrases were taken from the Google syntactic n-grams (Lin et al., 2012) by selecting the 100 most frequent phrases for each verb and manually filtering them to remove redundancies. This resulted in 10 phrases per verb. Examples for *ride* include *riding a horse*, *riding a skateboard*,

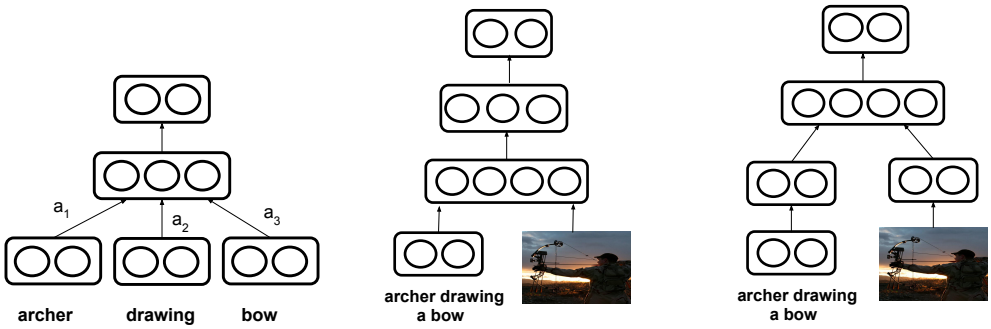
[1] https://en.wiktionary.org/

Query	catching a ball	archer drawing a bow
Verb	English: catching	English: drawing
	German: fangen	German: ziehen
	Spanish: coger	Spanish: tensar
Description	a baseball player is trying to catch a ball	A man is holding a microphone in his hand
Situation	ACTIVITY: tackling	ACTIVITY: aiming
	AGENT: football player	AGENT: man
	VICTIM: football player	ITEM: bow
	PLACE: football field	TARGET: None
		PLACE: outdoors

Figure 4.2: Example image from MultiSense annotated with German and Spanish verb translations. We also show the automatic image description generated using the Visual Sentinel model (Lu et al., 2017) and the situation predicted using imSitu (Yatskar et al., 2016).

and *riding a bicycle*. We also made sure that different uses of a verb are covered in the phrases, that is maintaining the diversity among uses of the verb. For every phrase we retrieved 150 Safe Search images from Google Image Search. We filtered the retrieved images using Amazon Mechanical Turk, where the crowd workers are also shown the query which was used to retrieve images. The crowd workers selected those images that are relevant to the phrase and depict the activity denoted by it. Annotators removed clip art and computer-generated images.

We employed native German and Spanish speakers to translate the verb phrases without visual context but were informed the usecase of translation. We assign the verbs from the query translation are then used verb annotations for the associated images. Overall pairwise agreement for image filtering task was 0.763. We followed this approach instead of using German and Spanish query translations to obtain new images because (1) image retrieval works better for English, i.e., returns a higher percentage of relevant images, (2) this approach ensures that the images are the same for all three



(a) Weighted embedding (b) Early multimodal fusion (c) Late multimodal fusion

Figure 4.3: (a) architecture for attention-based sentence embedding based only on text; (b) architecture for early multimodal fusion of text and image representations; (c) architecture for late multimodal fusion.

languages (if we used query translations for image retrieval we would get a different set of images for each language).

This resulted in 9,504 images in total covering 55 verbs with 154 and 136 unique translations in German and Spanish, respectively. The average dispersion of the images for all verbs in MultiSense is 0.298. The image dispersion measure is considered a substitute for the concreteness of a word (Kiela et al., 2014) and is calculated as:

$$d = \frac{2}{n(n-1)} \sum_{i<j\leq n} 1 - \cos(v_i, v_j) \tag{4.1}$$

where v_i is the visual representation of the image i. Higher dispersion indicates lower concreteness of concept or verb in our case. Image dispersion d of a verb v is defined as the average pairwise cosine distance between all the visual representations $v_1 \ldots v_n$ in the set of images for that verb in our dataset. Lower dispersion rate here indicates high concreteness i.e., good visual signal from our images. Please refer to Hartmann & Søgaard (2018) for analysis on image dispersion and its impact on using image representations in cross-lingual tasks such as learning cross-lingual word representations.

We divided MultiSense into 75% training, 10% validation and 15% test splits. Example images for the verb *blow* are shown in Table 4.2.

Sentence-level Translations We also collected a dataset of sentence-level translations for the English and German data. This is a new resource of 995 image descriptions that makes it possible to evaluate the verb sense disambiguation capabilities of multimodal MT models. We collected the dataset in four-steps: (1) crowdsource English descrip-

tions of the images using the gold-standard MultiSense verb as a prompt; (2) manually post-edit the English descriptions to ensure they contain the correct verb; (3) crowdsource German translations, given the English descriptions, the German gold-standard MultiSense verb, and the image; (4) manually post-edit the German translations to ensure they contain the correct verb.

Throughout the data collection process, we calculated the accuracy with which the correct verb occurred in the sentences. We PoS tagged and lemmatized the sentence tokens and gold-standard verbs to reduce problems with surface-form matching. A verb was counted as correct if it was marked with the coarse-grained PoS VERB and its lemmatized form was the same as that of the gold-standard verb in MultiSense for that image. The verb accuracy of the English crowdsourced data was 47.9%, with many workers ignoring the task instructions and not using the prompt. After post-editing by one native and one fluent English speaker verb accuracy was 97.7%. The remaining errors are due to the SpaCy lemmatizer not correcting lemmatizing *riding* due to PoS tagging errors. The accuracy of the crowdsourced German translations was 82.9%, and manual post-editing by one native speaker increased it to 96.7%.

In order to quantify the difficulty of translating MultiSense image descriptions, we calculated the type-based out-of-vocabulary (OOV) rate using calculated using the Moses `oov.pl` script after normalizing, tokenizing, and lowercasing the text in each language. MultiSense dataset has a 14.3% English OOV rate, and a 25.2% German OOV rate, with respect to the Multi30k training data. The Multi30k 2017 English test set has an 10.1% OOV rate, and the German data has a 19.2% OOV rate. Because of the higher OOV rate, and because of high verb ambiguity (by design), the MultiSense descriptions should be more difficult to translate than the standard Multi30k ones.

4.4 Verb Sense Disambiguation Modeling

The existing literature in traditional WSD, cross-lingual WSD, and visual sense disambiguation, shows that supervised models outperform unsupervised models, even if only a small training set is available (e.g., Gella et al., 2016). In this chapter, we will therefore focus on supervised methods for cross-lingual WSD using MultiSense. Our main question is whether multimodal supervision outperforms the use of textual or visual features alone. We build verb-specific models to identify the correct translation of an English verb in the target language, as well as verb-independent models that can generate translations for all the 55 verbs in MultiSense.

4.4.1 Visual Classifiers

Visual features extracted from images have shown to be useful in learning various tasks such as learning semantic representations of words (Lazaridou et al., 2015), bilingual lexicon learning (Bergsma & Van Durme, 2011; Kiela et al., 2015; Hartmann & Søgaard, 2018), human object interactions (Chao et al., 2015a) and visual sense disambiguation (Gella et al., 2016). To test whether visual features can also be used to identify correct verb translations, we fine-tune a convolution neural network (CNN) image classifier for each verb in MultiSense. Separate models are trained for each target language. Per word classifiers have been used for both textual cross-lingual and visual word sense disambiguation tasks (Lefever & Hoste, 2013; Gella et al., 2016). We test the significance of visual features using (1) a feed-forward neural network model with a single hidden layer, whose input is the visual features extracted from an object classifier; (2) aggregated representations of visual features extracted from the object classifier. Following the previous work of Kiela et al. (2015), we experimented with two ways of aggregating the visual features of an object classifier for given label, viz., CNN-Mean, the component-wise average, and CNN-Max, the component-wise maximum of all the images bearing the same label in the training set. We use the aggregated representations of the training images to compute the similarity with a test image and assign as verb translation the label with highest similarity.

4.4.2 Textual Classifiers

Except for a small subset for MT evaluation, the images in MultiSense do not have text associated with them. Hence we explore three ways of obtaining textual context: (1) the query phrase (Q) used to retrieve images from the web functions as textual context; (2) an automatic image description (D) is generated using a state-of-the-art system (Lu et al., 2017); (3) a situation (S), also known as a visual frame, is predicted using a situation recognition algorithm; the situation includes the activity depicted and the objects involved in performing it (Yatskar et al., 2016).

We present example images with generated textual context in Figure 4.2. For situations, we ignore the role labels (activity, agent, etc.) and only consider the predicted role fillers. For example, for the image in Figure 4.2, the situations prediction is *man aiming bow outdoors*. To have comparable models for all three types of context (queries, descriptions, and situations), we consider the textual context as a bag of words instead of as a sequence of words.

To disambiguate verbs based on textual context, we use a feed-forward neural network with single hidden layer that takes an average of word embeddings of dimension d as input to classify the verb translations. Instead of a simple average, we compute the weighted sum of the embeddings of the words in the context (shown in Figure 4.3(a)). This mechanism is similar to the global attention model of Luong et al. (2015), used to decode sentences in neural machine translation, and to the sentence embedding model of He et al. (2017). We also experimented with employing a recurrent neural network (RNN) to compute text-embeddings. However, we found that our simple average or weighted sum of the word embeddings models performed better than RNN based representation for text.

More formally, for the textual context c of each input image, a vector representation c_t is computed as the weighted sum of word embeddings e_{w_i}, where $i = 1, \ldots, n$ correspond to the indices of the words in the textual context: We padded all context to be of a fixed length n.

$$c_t = \sum_{i=1}^{n} a_i e_{w_i} \tag{4.2}$$

For each word w_i in the context, we compute a weight a_i using an attention model conditioned on the embedding of the word w_i as well as the overall embedding of the textual context e_c, i.e., the average of all the word embeddings in the context:

$$a_i = \frac{\exp(d_i)}{\sum_{j=1}^{n} \exp(d_j)} \tag{4.3}$$

$$d_i = e_{w_i}^T \cdot W_a \cdot e_c \tag{4.4}$$

A weight matrix $W_a^{|d \times d|}$, a mapping between the overall embedding of the textual context e_c and the word embedding e_{w_i} is learned during training:

$$e_c = \frac{1}{n} \sum_{i=1}^{n} e_{w_i} \tag{4.5}$$

With this attention model we aim to filter out words that do not contribute to the translation of the verb from textual context by down-weighting them using attention.

4.4.3 Multimodal Classifiers

In addition to using image representations and textual context representations separately, we also consider multimodal representations that integrate the visual and textual

features. In order to obtain such multimodal representations, we concatenate the CNN features extracted from image with the textual embeddings. We experiment with various ways of fusing textual and visual features. The simplest option is early multimodal fusion which involves concatenating textual and visual features as inputs and passing them through a feed-forward network, as illustrated in Figure 4.3(b). The late multimodal fusion involves a two-layer network that takes textual and visual inputs and passes them through separate feed-forward layers before concatenating them. For both early and late fusion, we experiment with a simple average of textual context embeddings, as well as with weighted average embeddings (as described in Section 4.4.2).

4.5 Verb Disambiguation Experiments

4.5.1 Experimental Setup

We used the pre-trained $d = 300$ dimensional word-embeddings available with the word2vec package. They were trained on a part of the Google News dataset of approximately 100 billion words (Mikolov et al., 2013a). To extract the visual features, we use the ResNet-34 architecture (He et al., 2016) pre-trained on the ImageNet object classification dataset.

All of our networks were trained using stochastic gradient descent with mini-batches of 16 samples. We chose the hidden layer dimensionality as $D = 128$. We used the state-of-the-art Visual Sentinel image description system (Lu et al., 2017) to generate image descriptions for all the images in MultiSense. To predict the situations, we used the neural-CRF based situation prediction model of Yatskar et al. (2016), fine-tuned on the ResNet-101 architecture.

To evaluate the proposed models, we compared against chance and majority label baselines. For visual features, we also compared against baselines using aggregated features, obtained using CNN-Max and CNN-Mean aggregation (see Section 4.4.1). We compare two different classification approaches for both languages: in the setting AVG, we build individual classifiers for each verb and report averaged scores across all 55 verbs in our dataset. In the setting ALL, we build a single classifier for each language for all the verbs in the dataset. In both the settings, we experiment with textual, visual, and multimodal features. For both textual and multimodal we use simple averaging of textual embeddings, as well as attention, i.e., the learned weighted average of the component embeddings. Finally, for our multimodal setup, we experimented with both

early fusion and late fusion.

German: Img: 928, Cha: 35.3, Maj: 51.6

Attn	Textual			Vis	F	Multimodal		
	D	S	Q			D	S	Q
N	75.0	79.8	**81.2**	91.4	E	**93.7**	93.7	92.5
Y	76.9	78.9	79.8	91.4	E	93.5	92.0	92.2
N					L	87.0	89.3	88.8
Y					L	89.4	89.2	89.3
Vis Baselines: CNN-Mean: 4.0, CNN-Mean: 43.6								

Spanish: Img: 929, Cha: 37.2, Maj: 57.0

Attn	Textual			Vis	F	Multimodal		
	D	S	Q			D	S	Q
N	63.5	**81.8**	81.1	91.7	E	**93.5**	93.3	93.3
Y	70.8	81.0	80.2	91.7	E	92.6	91.6	91.8
N					L	88.4	89.0	88.3
Y					L	88.9	89.3	88.8
Vis Baselines: CNN-Mean: 6.0, CNN-Max: 53.1								

Table 4.4: Accuracy of cross-lingual WSD on the **validation set** using all combinations of textual, visual, and multimodal models. Average accuracy is reported over separate models for each of the 55 verbs. Img: number of images; Cha: chance baseline; Maj: majority baseline; Attn: attention; F (fusion): early (E); or late (L) fusion; textual context: predicted image descriptions (D), predicted situations (S), gold queries (Q).

4.5.2 Results

An overview of the results of all our models on the validation set is given in Table 4.4 and Table 4.6. In Table 4.4 we present the results of AVG setting, where we have an individual classifier for each verb. In Table 4.6 we present results of ALL setting where a single model is learned for all 55 verbs jointly. In each table we present textual features using descriptions, situations and queries are input; visual features extracted from image; and multimodal features combining visual with textual features.

Lang	Setting	Cha	Maj	Text	Vis	MM
German	AVG	35.3	50.6	80.3	85.0	**88.3**
Spanish	AVG	37.2	55.2	79.4	85.7	**87.3**
German	ALL	0.65	2.8	49.1	52.1	**55.6**
Spanish	ALL	0.74	4.0	52.7	50.3	**56.0**

Table 4.5: Accuracy for cross-lingual WSD on the test set using best textual (Text), visual (Vis), and multimodal (MM) settings as determined on the validation set. Cha: chance baseline; Maj: majority baseline.

Attn	Textual			Vis	F	Multimodal		
	D	S	Q			D	S	Q
N	18.8	32.9	**51.0**	51.3	E	50.7	50.5	**54.3**
Y	23.6	30.9	48.7	51.3	E	51.0	51.0	51.5
N					L	47.4	47.0	46.4
Y					L	46.2	46.3	46.4
Vis Baselines: CNN-Mean: 0.0, CNN-Max: 0.1								

Attn	Textual			Vis	F	Multimodal		
	D	S	Q			D	S	Q
N	19.6	33.4	**54.0**	53.1	E	50.7	50.5	**54.3**
Y	22.8	32.1	53.5	53.1	E	49.4	49.8	50.3
N					L	46.5	46.0	48.4
Y					L	47.0	46.8	47.8
Vis Baselines: CNN-Mean: 0.0, CNN-Max: 4.0								

Table 4.6: Accuracy of cross-lingual WSD on the **validation set** using all combinations of textual, visual, and multimodal models. Accuracy is reported for a single model over all verbs. Img: number of images; Cha: chance baseline; Maj: majority baseline; Attn: attention; F (fusion): early (E); or late (L) fusion; textual context: predicted image descriptions (D), predicted situations (S), gold queries (Q).

These results show that multimodal classification models that integrate textual and

visual features outperform models based on visual or textual features alone in both of our settings (AVG and ALL). Furthermore, we found that integrating text and vision via early fusion, i.e., by concatenating visual features and textual embeddings at the input layer (as shown in Figure 4.3(b)) performed better than late fusion, where concatenation happens at the hidden layer (see Figure 4.3(c)). This result again holds for both AVG and ALL models.

Regarding the usefulness of attention-based weighting, we found that attention is useful only for certain types of textual information. In a text-only setting, attention improves results when the text is derived from image descriptions (D), but not when it is derived from situations (S), or queries (Q). It seems that the length of the textual context matters: descriptions are longer than situations or queries, and attention is able to down-weight the irrelevant words that occur in long textual contexts. The average length of descriptions, situations and queries in the training set is 6.30, 3.95, and 2.58, respectively (We did not remmove the stopwords for computing the embeddings, this is just removed to compute the length of the various textual inputs). Note that stop words were removed in all three cases. In a multimodal setting, the picture is more mixed: in some of the late-fusion models, attention yields a slight improvement in performance. However, in no case is attention part of the best performing model.

Finally, we can address the issue of which type of textual and visual information is most useful. In the text-only setting, we found that queries are most useful, followed by situations. Descriptions consistently under performed. This holds both for AVG and ALL models. For vision-only models, fine tuning substantially outperforms the CNN-Mean and CNN-Max baselines (we therefore do not report any multimodal experiments with these baselines). For multimodal models, all types of textual information perform in a broadly similar way. The overall winner is situations for German and descriptions for Spanish in the AVG setting, and queries for both languages in the ALL setting. This is an encouraging result, as it demonstrates that we can do without gold-standard text (the queries were manually provided), at least in the AVG setting.

We present the results on the test set in Table 4.5. We follow standard practice of evaluating all models on the validation set, select the best model per setting model per setting (AVG vs. ALL) and modality (textual, visual, multimodal) and then evaluate on the tests set. The test set results confirm the overall patterns we observed on the validation set: for unimodal models, visual features outperform textual features, but multimodal models outperform unimodal models. This result holds for both AVG and ALL settings, and for both German and Spanish.

Image	Context	German predictions	Spanish Predictions
	Q: **blowing** a balloon	blasen, **aufblasen**, steigen	**hinchar**, soplar, subir
	Visual features	**aufblasen**, drücken, drehen	**hinchar**, apretar, hacer, girar
	D: a woman holding a pink frisbee in her hand	tragen, schütteln, blasen	tapar, agitar, mirar
	D + Visual features	**aufblasen**, drücken, drehen	**hinchar**, apretar, hacer, girar
	Q: **serving** a volleyball	**spielen**, dienen, anschauen	**sacar**, servir, golpear
	Visual Features	**spielen**, spreizen, reiten	**sacar**, coger, abrir
	S: scoring basketball court basketball player	abblocken, **spielen**, treffen	taponar, **sacar**, golpear
	S + Visual Features	**spielen**, tragen, reiten	**sacar**, llevar, coger
	D: a person holding a cell phone in their hand	schütteln, drücken, schauen	ver, hablar, pasar el rato
	Visual Features	tragen, nachschauen, schütteln	**buscar**, mirar, pasar el rato
	Q: **looking** for directions	schauen, **suchen**, anschauen	mirar, **buscar**, subir
	Q + Visual Features	**suchen**, einpinseln, nachschauen	**buscar**, mirar, sonar

Table 4.7: Images with different contexts (textual, visual, or both) and the top three verb translation predictions from our textual and visual classifiers in the ALL setting. Q: query used to retrieve images; D: Image description generated by Lu et al. (2017); S: Situation predicted using Yatskar et al. (2016). The ground truth label for the image is in **bold**.

4.5.3 Discussion

We analyzed the outputs of our models in order to understand where multimodal features helped in identifying the correct verb translation and the cases where they failed. In Table 4.7, we show selected examples that illustrate how varying context (textual, visual or multimodal) affects the performance of verb translation. For every image we give the top three verbs predicted by our models in the ALL setting for both German and Spanish. For the first image of *blowing balloon*, the automatically generated description is not relevant to the image and thus results in an incorrect verb in both German and Spanish. When visual information is added to descriptions, the model is able to produce the correct verb. For the second image showing *serving a volleyball*, the top predicted verb using the situation context alone is incorrect. However, when visual information is added, the model predicts the correct label. These examples demonstrate that both the image description system and the situation prediction system often generate irrelevant text, whereas the CNN features extracted from the images tend to generalize well across verbs. Normally, words in isolation are not translated, a valid usecase for such scenario is translation of tags or hashtags with visual information.

4.6 Constrained Decoding

In real-world scenarios, we might have access to additional information of the data point at inference time that will help improve the performance of the model. For example, while doing domain adaptation in Machine Translation if the domain of the input is known at prediction time with access to domain terminology or a named entity recognizer we can ensure that specific domain specific terms are present in system outputs. Similar to the problem we address here text generation using multi-modal input, we might be able to detect actions, objects in images that could be mentioned in the output using auxiliary models. The goal of constraint decoding models is to generate best output using extra information that is provided to the model at inference time which otherwise model will not have access to.

For machine translation models at inference time model generates the sequence $\hat{y} = \{y_0, y_1, y_2..y_T\}$ of length T for a given input sequence x, that has the maximum probability parameterized by a model θ:

$$\hat{y} = argmax_{y \in Y} p_\theta(y|x)$$

In constrained decoding, an additional input of constraints $\{c_0 \dots c_n\}$ is provided and the model that tries to find the optimal sequence which includes the set of given lexical constraints. A standard approach is to generate the output sequence from beginning to end, conditioning the output at each timestep t upon the input x and the previously generated symbols $\{y_0, \dots y_{t-1}\}$. Often beam search is employed to avoid the risk of making locally optimal decisions while decoding and to avoid exhaustive exploration of the output sequences (Och & Ney, 2004). Post & Vilar (2018) proposed a dynamic vairant of lexically constrained grid beam search (GBS) algorithm that can constrain the search space to outputs which contain one or more pre-specified sub-sequences or constraints. They use the model's distribution to lexical constraints correctly as well as to generate the parts of the output which are not covered by the constraints. Please review Hokamp & Liu (2017) and Post & Vilar (2018) for algorithmic details of constraint decoding.

4.7 Machine Translation Experiments

We now evaluate the utility of our verb sense disambiguation model for the challenging downstream task of multimodal machine translation. We address this as a enhancement of text-only machine translation model which do not have access to images with translation data at training time where as sentences at test time do have visual information or image data. We first build text-only machine translation model (our baseline) then add verb translation as an extra input to decoder. We use a lexically constrained decoder which is a modification to a standard decoder with beam search that allows to specify words that must appear in the target language.

We conduct this evaluation of on the image descriptions test set of MultiSense (see Sentence-level Translations in Section 4.3). We calculated BLEU (Papineni et al., 2002) and Meteor scores (Denkowski & Lavie, 2014) between the MultiSense reference description and the out of the translation model. We also evaluate the verb prediction accuracy of the MT output against the gold standard verb annotation.

4.7.1 Models

Our baseline is a single-layer attention-based text-only neural machine translation model (Hieber et al., 2017) trained on the 29,000 sentence English-German parallel text in Multi30k (Elliott et al., 2016). We preprocessed the data by normalizing punctuation,

	Meteor	BLEU	VAcc
Baseline NMT	38.6	17.8	22.9
+ Predicted	40.0	18.5	49.5
+ Oracle	40.4	19.1	77.7
Caglayan et al. (2017)	46.1	25.8	29.3
Helcl & Libovický (2017)	42.5	22.3	25.1

Table 4.8: MT results: Meteor and BLEU are standard text-similarity metrics, and verb accuracy (VAcc) counts how often the model proposal contains the gold standard German verb.

tokenizing the text, and the lowercasing it. We then learned a joint byte-pair-encoded vocabulary with 10,000 merge operations to reduce sparsity (Sennrich et al., 2016).

Our model uses the prediction of the verb disambiguation model as an additional input to a unimodal translation model. More specifically, we use the WSD verb prediction as a constraint on the lexically-constrained decoder of the MT model (Post & Vilar, 2018). We compare the performance of this setup against two state-of-the-art multimodal English–German translation systems: Caglayan et al. (2017), where the target language word embeddings are modulated by an element-wise multiplication with a learned transformation of the visual data; and Helcl & Libovický (2017), a double attention model that learns to selectively attend to a combination of the source language and the visual data.

4.7.2 Results

Table 4.8 shows the results of the translation experiment. Overall, the Meteor scores are much lower than on the Multi30k test sets, where the state-of-the-art single model scores 51.6 Meteor points compared to 46.1 Meteor we obtained. This gap is most likely due to the higher out-of-vocabulary rate in MultiSense (see Section 4.3). Using Predicted verb constraints outperforms the text-only translation baseline by 1.3 Meteor points. Furthermore, the translation output of our model contains the expected German verb 27% more often than the unimodal baseline model. These results show that a multimodal verb sense disambiguation model can indeed improve translation quality in a multimodal setting.

We also calculated the upper bound of our model by using the gold standard Ger-

Source	A woman smiles as she **brushes** her long, dark hair.
Ref	Eine Frau lächelt während sie sich ihre dunklen langen Haare **bürstet** .
Baseline	eine frau lächelt , als sie ihren langen und dunklen haaren **putzt** .
+WSD	+(bürsten): eine frau lächelt , als sie ihr lange , dunklen haaren **bürsten** .

Source	A large herd of sheep is **blocking** the road.
Ref	Eine große Herde Schafe **blockiert** die Straße .
Baseline	eine große herde schafe **kriecht** die straße entlang .
+WSD	+(blockieren): eine große herde schafe **blockieren** die straße .

Table 4.9: Examples where our WSD prediction input improves translations. Wrong verbs in baseline translations are shown in **red.**

man verb as the lexical constraint. In this oracle experiment we observed a further 0.7 Meteor point improvement over our best model, and a further 27% improvement in verb accuracy. This shows two things: (1) there are further improvements to be gained from improving the verb disambiguation model, and (2) the OOV rate in German means that we cannot achieve perfect verb accuracy.

4.8 Discussion and Conclusions

We introduced the task of cross-lingual visual sense disambiguation for verbs: given a verb and an image as visual context, identify the correct translation of the verb in a target language. We developed the new MultiSense dataset, in which 9,504 images are annotated with one of 55 verbs and its translations in Spanish and German. We proposed a series of supervised cross-lingual visual sense disambiguation models and showed that multimodal models that fuse textual information (generated by image description or situation recognition systems) with visual features outperform unimodal models.

Cross-lingual WSD is an interesting problem in its own right, but it also has a clear application in machine translation. Determining the correct sense of a verb is important for high quality translation output, and sometimes text-only translation systems fail when the correct translation would be obvious from visual information (see Figure 4.1). To show that cross-lingual visual sense disambiguation can improve the performance of translation systems, we annotated a part of our MultiSense dataset with English image descriptions and their German translations.

There are two existing multimodal translation evaluation sets with ambiguous words: the Ambiguous COCO dataset Elliott et al. (2017) contains sentences that are "possibly ambiguous", and the Multimodal Lexical Translation dataset is restricted to predicting single words instead of full sentences Lala & Specia (2018). MultiSense contains sentences that are known to have ambiguities, and it allows for sentence-level and verb prediction evaluation. Here, we use the verbs predicted by our visual sense disambiguation model to constrain the output of a neural translation system and demonstrate a clear improvement in Meteor, BLEU, and verb accuracy over a text-only baseline.

In this work, we only aim to address ambiguity across languages for visual verbs. However, this do not address all senses of verbs as in realistic scenario texts presented for Machine Translation could have non-visual senses as well. This would a direction for future work to explore the other information from non-visual senses to improve disambiguation across languages in Machine Translation.

Chapter 5

Image Pivoting for Learning Multilingual Multimodal Representations

In this chapter, we propose a model to learn multimodal multilingual representations for mapping images and sentences into common embedding space. Although there exist models which learn to map sentences and images into a common embedding space in order to be able to retrieve one from the other, most of the existing models focused on single language especially English. The novelty of this work is in mapping sentences from multiple languages and images into a common space and evaluating the usefulness of the second language in multimodal search.

The main focus of this work is advancing multilingual versions of image search and image understanding. Given an image and its descriptions in two different languages which need not be parallel, our proposed model learns a common representation for images and their descriptions in two different languages by considering the image as a pivot between two languages. We propose a new pairwise ranking loss function which can handle both symmetric and asymmetric similarity between the two modalities. We evaluate our models on image-description ranking for German and English and compare performance with existing state-of-the-art methods on image-description ranking. Additionally, we also evaluate our models on semantic textual similarity of image descriptions in English and for cross-lingual image description generation tasks.

5.1 Motivation

There has been a significant amount of research in joint modeling of texts and images. Examples include text-based image retrieval, image description and visual question answering. In the last few years an increasing number of large image description datasets has become available (Hodosh et al., 2013; Young et al., 2014c; Lin et al., 2014) which have been used for joint modeling of text and images.

A large number of systems have been proposed to handle the image description task as a generation problem (Bernardi et al., 2016; Mao et al., 2015; Vinyals et al., 2015; Fang et al., 2015a). Prior to that, there has also been a great deal of work on sentence-based image search or cross-modal retrieval where the objective is to learn a joint space for images and text (Hodosh et al., 2013; Frome et al., 2013; Karpathy et al., 2014; Kiros et al., 2014; Socher et al., 2014; Donahue et al., 2015; Yan & Mikolajczyk, 2015b).

Previous work on image description generation or learning a joint space for images and text has mostly focused on English due to the availability of English datasets. Recently there have been attempts to create image descriptions and models for other languages (Funaki & Nakayama, 2015b; Elliott et al., 2016; Rajendran et al., 2016; Miyazaki & Shimizu, 2016) (See Table 5.1 for statistics of existing multilingual image description corpora). However, they exist only for few languages such as German: (Grubinger et al., 2006; Elliott et al., 2016; Rajendran et al., 2016; Elliott et al., 2016), Japanese: (Funaki & Nakayama, 2015b; Miyazaki & Shimizu, 2016) and are small scale compared to existing English image description datasets (Lin et al., 2014).

Most of these datasets have been used to study description generation problem or for other tasks such as multimodal machine translation. It is not entirely clear how text-query based image-retrieval model models perform for queries in languages other than English. Querying an online image-search engine such as Google with a simple sentence query in English and its translation in German retrieves very different images. English queries tend to retrieve relevant images to query and German query retrieves worse results (shown in Figure 5.1). This highlights one of the key issues that available multimodal applications for other languages are far behind English.

Most work on learning a joint space for images and their descriptions is based on Canonical Correlation Analysis (CCA) or neural variants of CCA over representations of image and its descriptions (Hodosh et al., 2013; Andrew et al., 2013a; Yan & Mikolajczyk, 2015b; Gong et al., 2014a; Chandar et al., 2016). Besides CCA, a few others

A dog is chasing a mouse



Ein Hund jagt eine Maus



Figure 5.1: Google Image Search results for identical query in English and German. Retrieved results for English query results are more relevant to the query than that of its German translation.

learn a visual-semantic or multimodal embedding space of image descriptions and representations by optimizing a ranking cost function (Kiros et al., 2014; Socher et al., 2014; Ma et al., 2015; Vendrov et al., 2016) or by aligning image regions (objects) and segments of the description (Karpathy et al., 2014; Plummer et al., 2015) in a common space. Recently Lin & Parikh (2016) have leveraged visual question answering models to encode images and descriptions into the same space.

However, all of this work is targeted at monolingual descriptions (for English language only), i.e., mapping images and descriptions in a single language onto a joint embedding space. In this chapter, we explore the idea of of pivoting or bridging languages via visual information i.e., image. The idea of bridging is not new and is well explored for various natural language processing tasks such as machine translation (Wu & Wang, 2007; Firat et al., 2016) and to learn multilingual multimodal represen-

Corpus	Source	L1	L2	#Imgs	#L1	#L2	Para
UIUC-Pascal-JP (Funaki & Nakayama, 2015b)	Pascal	En	Jp	1k	5k	5k	Y
YJ-Captions (Miyazaki & Shimizu, 2016)	MSCOCO	En	Jp	26k	132k	131k	N
De-COCO (Hitschler et al., 2016)	MSCOCO	En	De	1k	1k	1k	Y
BridgeCorr (Rajendran et al., 2016)	MSCOCO	En	Fr	1k	5k	5k	Y
BridgeCorr (Rajendran et al., 2016)	MSCOCO	En	De	1k	5k	5k	Y
Flickr8k-CN (Li et al., 2016)	Flickr8k	En	Cn	8k	40k	40k	N
Flickr8k-CN (Li et al., 2016)	Flickr8k	En	Cn	1k	5k	5k	Y
Multi30k (Elliott et al., 2016)	Flickr30k	En	De	30k	30k	30k	Y
Multi30k (Elliott et al., 2016)	Flickr30k	En	De	30k	150k	150k	N
IAPR-TC (Grubinger et al., 2006)	IAPRTC	En	De	20k	35k	35k	N

Table 5.1: Statistics of multilingual image description corpora; Para: If the image descriptions are parallel corpora

tations (Rajendran et al., 2016; Calixto et al., 2017b). However, all of that work has used language as bridge to connect other pair of languages or between language and visual information. Unlike previous work, we learn representations where we use visual information as bridge between two languages.

Related to our work Calixto et al. (2017b) proposed a model for creating multilingual multimodal embeddings. Our work is different from theirs in that we choose the image as the pivot and use a different similarity function. Our proposed model is a single model for learning representations of images and multiple languages, whereas their model is language-specific. Rajendran et al. (2016) propose a model to learn common representations between M views and assume there is parallel data available between a pivot view and the remaining $M-1$ views. Their multimodal experiments are based on English as the pivot and use large parallel corpora available between languages to learn their representations.

In this chapter, we learn multimodal representations in multiple languages, i.e., our model yields a joint space for images and text in multiple languages using the image as a pivot between languages. We propose a new objective function in a multitask learning setting and jointly optimize the mappings between images and text in two different languages.

English descriptions	1) Two professional men's soccer players playing soccer. 2) Two men playing soccer on a field. 3) Two soccer players on a green field play with a soccer ball. 4) Two men, one wearing red uniform with white stripes and the other a white uniform are playing soccer. 5) Two soccer players, one wearing red and one wearing white, are competing to kick a soccer ball across a soccer field.
German descriptions	1) Zwei männer kämpfen einen fussball. 2) Szenen eines fussballspieles. 3) Zwei Männer spielen fußball. 4) Zwei Fussballer zweier Mannschaften jagen auf dem Spielfeld im Freien dem Ball hinterher. 5) Ein Fußballspieler im rot-weißen Trikot spielt den Ball, während ein anderer im blau-weißen Trikot von rechts angelaufen kommt.

Figure 5.2: Example annotation of an image in Multi30k description corpus.

Corpus	Language	Images	Sentences	Types	Tokens	#AvgL	Singletons
Translations	English	31,014	31,014	11,420	357,172	11.9	5,073
	German			19,397	333, 833	11.1	11,285
Descriptions	English	31,014	155,070	22,815	1,841,159	12.3	9,230
	German			46,138	1,434,998	9.6	26,510

Table 5.2: Corpus statistics of Multi30k dataset. #AvgL: Average length of description for each language. Singletons: Number of words that are only observed once in the corpus. Number of singletons in German descriptions are around 58% of the vocabulary.

5.2 Dataset

We experiment with the Multi30k dataset, a multilingual extension of Flickr30k corpus (Young et al., 2014c) consisting of English and German image descriptions (Elliott et al., 2016). The Multi30K dataset has 29000, 1014 and 1000 images in the train, validation and test splits respectively, and contains two types of multilingual annotations: (i) Translations: a corpus of one English description per image and its translation into German; and (ii) Descriptions: a corpus of five independently collected English and German descriptions per image (see Figure 5.2 for an example annotation).

We use the independently collected English and German descriptions to train our models. Note that these descriptions are not translations of each other, i.e., they are not parallel, although they describe the same image. English and German descriptions have 22.8k and 46.1k word types respectively i.e., number of word-types for German are more than double the number of word-types for English (see Table 5.2 for more corpus-level statistics). And the average number of words in English and German descriptions are 12.3 and 9.6.

Overall English descriptions has smaller vocabulary (22, 815 vs. 46,138) but the average length is higher than German descriptions. Also, the number of singletons in German descriptions is much higher than in the English descriptions (58% for German and 40% for English). The main reason for this could be word compounding observed in German as well as richer morphological variation. The English description corpus of Multi30k has been used extensively to learn and evaluate multimodal representations for image-description search systems (Kiros et al., 2014) and for building automatic image description systems (Karpathy & Li, 2015b). And recently, the translations in Multi30k dataset has been used for building multimodal machine translation systems.

5.3 Problem Formulation

Given an image i and its descriptions c_1 and c_2 in two different languages our aim is to learn a model which maps i, c_1 and c_2 onto the same common space $\mathbb{R}^N$ (where N is the dimensionality of the embedding space) such that the image and its gold-standard descriptions in both languages are mapped close to each other. Our model consists of the embedding functions f_i and f_c to encode images and descriptions and a scoring function S to compute the similarity between a description–image pair.

In the following we describe two models: (i) the PIVOT model that uses the image as pivot between the description in both the languages (shown in Figure 5.3) ; (ii) the PARALLEL model that further forces the image descriptions in both languages to be closer to each other in the joint space (as shown in Figure 5.4). . We build two variants of PIVOT and PARALLEL with different similarity functions S to learn the joint space.

5.3.1 Multilingual Multimodal Representation Models

In both PIVOT and PARALLEL models we have two main components textual component to encode description and a visual component to encode the image. For the visual component, we use a deep convolutional neural network architecture (CNN) to represent the image i denoted by $f_i(i) = W_i \cdot CNN(i)$ where W_i is a learned weight matrix and $CNN(i)$ is the image vector representation. The dimensions of the learned weight matrix W_i is $N \times D$, N being the dimensionality of embedding space and D denotes dimensionality of the CNN representation of the image, the activations of the last fully connected layer (fc7) of the CNN architecture.

In our textual component for each language we define a recurrent neural network encoder $f_c(c_k) = GRU(c_k)$ with gated recurrent units (GRU) activations to encode the description c_k in language k into a N dimensional vector. We have a separate gated recurrent unit encoder for each language k. Let $c_k = \{w_1^k, w_2^k, w_3^k, \ldots, w_{M_k}^k\}$ denote a description composed of M_k words in language k. A gated recurrent unit reads words from left to right and generates a sequence of recurrent annotation vectors $(h_1^k, h_2^k, h_3^k, \ldots, h_{M_k}^k)$ each with dimensionality of N. For a given description c_k, we use the last annotation vector $h_{M_k}^k$ as the description representation, henceforth $GRU(c_k)$.

In PIVOT, we use monolingual corpora from multiple languages of sentences aligned with images to learn the joint space. The intuition of this model is that an image is a universal representation across all languages, and if we constrain a sentence representation to be closer to images, sentences in different languages may also come closer.

Accordingly we design a loss function as follows:

$$loss_{pivot} = \sum_{k} \Bigg[\sum_{(c_k,i)} \Bigg(\sum_{c'_k} \max\{0, \alpha - S(\vec{c_k},\vec{i}) + S(\vec{c'_k},\vec{i})\} + \sum_{i'} \max\{0, \alpha - S(\vec{c_k},\vec{i}) + S(\vec{c_k},\vec{i'})\} \Bigg) \Bigg] \tag{5.1}$$

where k stands for each language.

This loss function encourages the similarity $S(\vec{c_k},\vec{i})$ between gold-standard description c_k and image i to be greater than any other irrelevant description c'_k by a margin α (a hyper-parameter we tune).

A similar loss function is useful for learning multimodal embeddings in a single language (Kiros et al., 2014). For each minibatch, we obtain invalid or contrastive descriptions by selecting descriptions of other images except the current image of interest and vice-versa.

In PARALLEL, in addition to making an image similar to a description, we make multiple descriptions of the same image in different languages similar to each other, based on the assumption that these descriptions, although not parallel, share some commonalities. Accordingly we enhance the previous loss function with an additional term:

$$\begin{aligned} loss_{para} = \sum_{k} \Bigg[\sum_{(c_k,i)} \Bigg(\sum_{c'_k} \max\{0, \alpha - S(\vec{c_k},\vec{i}) + S(\vec{c'_k},\vec{i})\} \\ + \sum_{i'} \max\{0, \alpha - S(\vec{c_k},\vec{i}) + S(\vec{c_k},\vec{i'})\} \Bigg) \Bigg] + \\ \sum_{(c_1,c_2)} \Bigg(\sum_{c'_1} \max\{0, \alpha - S(\vec{c_1},\vec{c_2}) \\ + S(\vec{c'_1},\vec{c_2})\} + \sum_{c'_2} \max\{0, \alpha - S(\vec{c_1},\vec{c_2}) + S(\vec{c_1},\vec{c'_2})\} \Bigg) \end{aligned} \tag{5.2}$$

Note that we are iterating over all pairs of descriptions (c_1, c_2), and maximizing the similarity between descriptions of the same image and at the same time minimizing the similarity between descriptions of different images.

Most approaches that project multiple views of data into a joint space are based on symmetric scoring function such as cosine similarity that maps semantically similar data points close-by in the embedding space (Hodosh et al., 2013; Kiros et al., 2014; Socher et al., 2014).

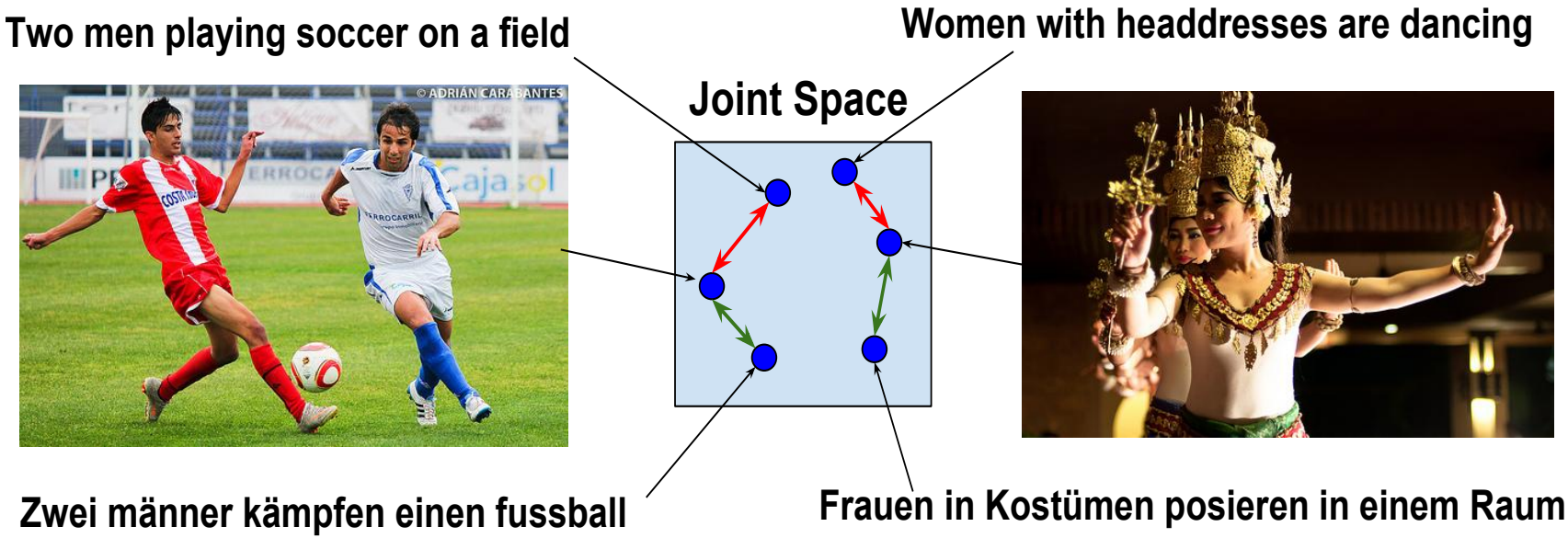


Figure 5.3: Our multilingual multimodal PIVOT model with image as pivot between two languages.

We learn models using two similarity functions: symmetric and asymmetric. For the former we use cosine similarity and for the latter we use the metric of (Vendrov et al., 2016) which is useful for learning embeddings that maintain an order, e.g., dog and cat are more closer to pet than animal while being distinct. Such ordering is shown to be useful in building effective multimodal space of images and texts. An analogy in our setting would be two descriptions of an image are closer to the image while at the same time preserving the identity of each (which is useful when sentences describe two different aspects of the image). The similarity metric is defined as:

$$S(a,b) = -||max(0, b-a)||^2 \tag{5.3}$$

where a and b are embeddings of image and description.

We call the symmetric similarity variants of our models as PIVOT-SYM and PARALLEL-SYM, and the asymmetric variants PIVOT-ASYM and PARALLEL-ASYM.

5.3.2 Baseline Models

VSE and OE: As baselines we use monolingual models, i.e., models trained on each language separately. Specifically, we use Visual Semantic Embeddings (VSE) of Kiros et al. (2014) and Order Embeddings (OE) of Vendrov et al. (2016). Visual Semantic embeddings model uses pairwise ranking loss function (shown in Equation 5.4, k represents the respective language.) that encourages ground-truth caption-image pairs to be closer to each other in the joint space. Both Visual Semantic Embeddings and Order Embeddings models are monolingual models that is trained separately for each language i.e., a separate model for each language trained using image-sentence pairs in respective language and the objective function below. Order embeddings model uses same pairwise ranking loss function and replaces symmetric similarity measure(S) with asymmetric order-violation penalty that encourages ground-truth caption-image pairs to be greater than that for all other contrastive caption-image pairs by a margin α:

$$\begin{aligned} loss_{baseline} = \sum_{(c_k,i)} \Bigg(& \sum_{c'_k} \max\{0, \alpha - S(\vec{c_k}, \vec{i}) + S(\vec{c'_k}, \vec{i})\} \\ & + \sum_{i'} \max\{0, \alpha - S(\vec{c_k}, \vec{i}) + S(\vec{c_k}, \vec{i'})\} \Bigg) \end{aligned} \tag{5.4}$$

where (c_k, i) denotes ground truth caption-image pair for language k, c'_k goes over all captions that do not describe image i in a single mini-batch and i' goes over all images that are not described by c in a mini-batch.

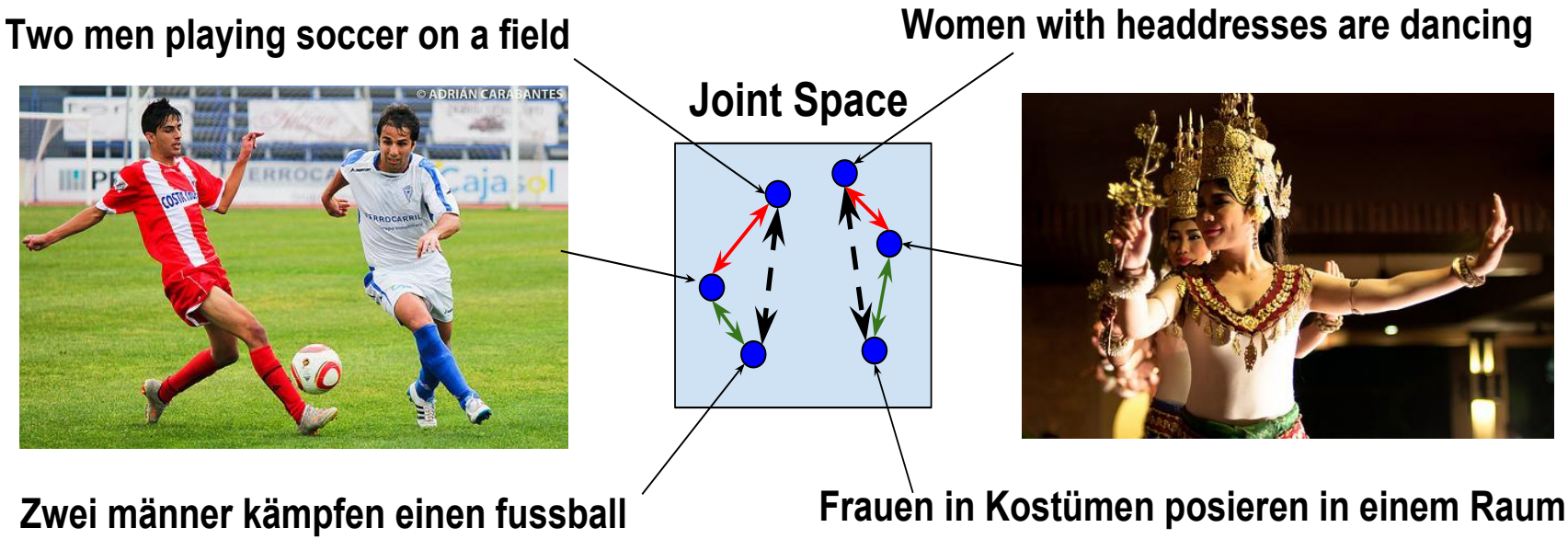


Figure 5.4: Our multilingual multimodal PARALLEL model with image as pivot and enforcing ground truth image descriptions in both languages to be closer to each other in the joint space.

Google: Additional to monolingual models we also combine them with state-of-the-art neural machine translation models to query the captions or images. If the query is in German and we have an English-Image (monolingual) model, we use a state-of-the-art neural machine translation model to translate German queries to English and vice-versa when the query is in English and the model is in German and report results. We use the online Google translate system available via its API to get the translations of English-German and German-English queries Wu et al. (2016). This would shed some light on applicability of machine translation models in multimodal tasks.

To measure the difficulty of this task on Multi30k dataset, we also include a random baseline. Additionally, we include a model with simpler text-representation that simply averages the embeddings of words in the description to create the sentence representation. For this we use the Average model studied by Hodosh (2015) which uses the average of Glove embeddings (Pennington et al., 2014) across all tokens to create sentence representation. This average sentence representation is used as input to a fully connected layer to learn text representation and uses a loss function identical to the Visual semantic embeddings model to learn joint space of text and images.

5.3.3 Comparison Systems for English Image-Description ranking

Our goal is to learn representations for multiple language and images together and evaluate the usefulness of the signal form other-language for image-description ranking. To answer this question, we compare our method with other related methods which use English image-description pairs to learn a joint embedded space. Most of the work on matching image and sentences have used convolutional neural networks occasionally with ground truth or predicted bounding box information of objects or entities present in the image (Plummer et al., 2015).

However, a large body of related work addressed how they encode the description and the matching of image and text representations (Ma et al., 2015). Here, we compare with three different such models that use a recurrent neural network, convolutional neural network and a Fisher vector pooling of word embeddings to generate sentence encoding respectively.

BRNN: Karpathy & Li (2015a) learns a model that is based on the insight that words or phrases in sentences written by people make references to a particular region of the image (for example objects, attributes, scenes) etc. For this they rely on a object detection system (Girshick et al., 2014) to identify the top N key regions in the image and encode each region using a CNN pre-trained on ImageNet and map them into a lower dimensional space h. They use a Bidirectional Recurrent Neural Network (BRNN) to transform each descriptions or a sequence of tokens into a h dimensional vector. They formulate a scoring function that maximizes the score for image-sentence if there is high matching between regions in the image and words in the sentence. They show that relying inferred alignments between words/phrases and regions in the image leads to improvement in image-sentence matching and the final ranking performance. Similar to many other methods they use max-margin loss function that encourages ground truth image-sentence pairs to have higher matching score than misaligned pairs.

m-CNN$_{st}$: Multimodal Convolutional neural network (m-CNN) is a framework proposed by Ma et al. (2015) that exploits image representation and word compositions and the matching relations between the two modalities. Unlike previous work on encoding descriptions with RNN, they propose matching CNN to compose different semantic fragments from words in the description and learn inter-modal relations between image and the semantic fragments at multiple levels.

For this, after encoding the image they produce a joint representation for image and each word or sequence of words in the description and pass the joint representation through a multilayer perceptron that produces the final matching scores between image and sentence. They experiment with mapping between images and words, phrases or sentences. They use a multimodal CNN that composes the whole sentence using matching CNN (m-CNN$_{st}$) that consists of three sequential layers of convolution and pooling and represent the whole sentence as a feature vector. Ma et al. (2015) also showed that ensemble models of word, phrase and sentence based matching CNN's perform slightly better than individual models alone.

CCA-FV: Klein et al. (2014) propose a new variant of Fisher vectors in combination with canonical correlational analysis to match texts and images. Fisher vectors are an advanced kernel based pooling technique that has been widely used for many computer vision applications such as image classification. In the proposed model each description is mapped to a set of vectors which are obtained via embeddings of individual words

in the description.

These set of vectors are combined to form a single vector based on concatenation of gradients of the log-likelihood of the individual vectors with respect to the parameters of Gaussian mixture model. They use word2vec embeddings (Mikolov et al., 2013a)to get the embeddings of individual words in the description and a convolution neural network pre-trained on ImageNet to encode the image and a canonical correlational analysis algorithm to match images and text representations.

5.4 Experiments and Results

We test our model on the tasks of image-description ranking. We work with each language separately. Since we learn embeddings for images and languages in the same semantic space, our hope is that the training data for each modality or language acts complementary data for the another modality or language, and thus helps us learn better embeddings.

5.4.1 Experiment Setup

We sampled minibatches of size 64 images and their descriptions, and drew all negative samples from the minibatch which result in 63 images for each description. We trained using the Adam optimizer with learning rate 0.001, and early stopping on the validation set. Following Vendrov et al. (2016) we set the dimensionality of the embedding space and the GRU hidden layer N to 1024 for both English and German.

We set the dimensionality of the learned word embeddings to 300 for both languages, and the margin α to 0.05 and 0.2, respectively, to learn asymmetric and symmetric similarity-based embeddings. We constrain the embeddings of descriptions and images to have non-negative entries when using asymmetric similarity by taking their absolute value. We keep all hyperparameters constant across all models. We used the L2 norm to mitigate over-fitting following prior work of Kiros et al. (2014). To pre-process data we tokenise and truecase both English and German descriptions using the Moses Decoder scripts.[1]

[1] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

System	Text to Image				Image to Text			
	Recall@1	Recall@5	Recall@10	Mr	Recall@1	Recall@5	Recall@10	Mr
Random	0.1	0.5	1.0	500	0.1	0.6	1.1	631
VSE (Kiros et al., 2014)	23.3	53.6	65.8	5	31.6	60.4	72.7	3
Google (En-De) + De-VSE	18.6	42.7	54.6	8	26.3	54.2	66.6	4
PIVOT-SYM	23.5	53.4	65.8	5	31.6	61.2	73.8	3
PARALLEL-SYM	24.7	53.9	65.7	5	31.7	62.4	74.1	3
OE (Vendrov et al., 2016)	25.8	**56.5**	67.8	4	**34.8**	**63.7**	74.8	3
Google (En-De) + De-OE	19.6	44.1	56.0	7	23.2	52.2	66.2	5
PIVOT-ASYM	26.2	**56.4**	**68.4**	4	33.8	62.8	**75.2**	3
PARALLEL-ASYM	**27.1**	**56.2**	66.9	4	31.5	61.4	74.7	3

Table 5.3: Image-description ranking results of English on Flickr30k test data. Both VSE and OE are monolingual models trained separately on each language. Google refers to the Google Translate model on test language to translate it to the language of the model trained.

5.4.2 Visual Feature Representation

To extract image features, we used a convolutional neural network model trained on 1.2M images of 1000 class ILSVRC 2012 object classification dataset, a subset of ImageNet (Russakovsky et al., 2015). Specifically, we used VGG 19-layer convolution neural network architecture and extracted the activations of the last fully connected layer to obtain features for all images in the dataset (Simonyan & Zisserman, 2014).

We use average features from 10 crops of the re-scaled images. We rescale images so that the smallest side is 256 pixels wide, we take 224 $\times$ 224 crops from the corners, center, and their horizontal reflections to get 10 crops for the image. Using average 10-cropped features has improvements in image description generation and learning joint multimodal space (Vendrov et al., 2016).

5.4.3 Image-Description Ranking Results

To evaluate the multimodal multilingual embeddings, we report results on an image-description ranking task. Given a query in the form of a description or an image, the task its to retrieve all images or descriptions sorted based on the relevance. We use the standard ranking evaluation metrics of recall at position k (Recall@K) and median rank (Mr) to evaluate our models. Recall (R@k) measures the mean number of captions for which the ground truth image is ranked within the top-k retrieved results (and vice-versa for images) and higher the recall better the model.

The Median rank (Mr) measures the median rank of the closest ground truth image or caption from the ranked list. Lower median rank implies better model. We report results for both English and German descriptions. Note that we have one single model for both languages.

In Tables 5.3 and 5.5 we present the ranking results of the baseline models of random, Kiros et al. (2014) and Vendrov et al. (2016) and our proposed PIVOT and PARALLEL models. We do not compare our image-description ranking results with Calixto et al. (2017b) since they report results on half of the validation set of Multi30k whereas our results are on the publicly available test set of Multi30k.

System	VF	Text to Image				Image to Text			
		Recall@1	Recall@5	Recall@10	Mr	Recall@1	Recall@5	Recall@10	Mr
Random	–	0.1	0.5	1.0	500	0.1	0.6	1.1	631
Glove-Average (Hodosh, 2015)	VGG16	19.1	47.6	61.0	5	26.9	56.8	69.2	3
BRNN (Karpathy & Li, 2015a)	VGG16	15.2	37.7	50.5	9	22.2	48.2	61.4	5
FV-CNN (Klein et al., 2014)	VGG19	24.4	52.1	65.6	5	34.4	61.0	72.7	3
m-CNN$_{st}$(Ma et al., 2015)	VGG19	19.7	48.4	62.3	6	27.0	56.4	70.1	4
OE (Vendrov et al., 2016)	VGG-19	25.8	**56.5**	67.8	4	**34.8**	**63.7**	74.8	3
PIVOT-ASYM	VGG-19	26.2	**56.4**	**68.4**	4	33.8	62.8	**75.2**	3
PARALLEL-ASYM	VGG-19	**27.1**	56.2	66.9	4	31.5	61.4	74.7	3

Table 5.4: Our best Image-description ranking results of English on Flickr30k test data compared to other state-of-the-art models.

System	Text to Image				Image to Text			
	Recall@1	Recall@5	Recall@10	Mr	Recall@1	Recall@5	Recall@10	Mr
Random	0.1	0.5	1.0	500	0.1	0.6	1.1	631
VSE (Kiros et al., 2014)	20.3	47.2	60.1	6	29.3	58.1	71.8	4
Google (De-En) + En-VSE	18.8	44.8	58.4	7	26.4	54.7	69.1	4
PIVOT-SYM	20.3	46.4	59.2	6	26.9	56.6	70.0	4
PARALLEL-SYM	20.9	46.9	59.3	6	28.2	57.7	71.3	4
OE (Vendrov et al., 2016)	21.0	48.5	60.4	6	26.8	57.5	70.9	4
Google (De-En)+ En-OE	21.1	48.1	60.7	6	27.3	59.7	71.7	4
PIVOT-ASYM	**22.5**	49.3	61.7	6	28.2	**61.9**	**73.4**	**3**
PARALLEL-ASYM	21.8	**50.5**	**62.3**	**5**	**30.2**	60.4	72.8	**3**

Table 5.5: Image-description ranking results of German on Flickr30k test data. Both VSE and OE are monolingual models trained separately on each language.

For English, PIVOT with asymmetric similarity is either competitive or better than monolingual models and symmetric similarity, especially in the R@10 category it obtains state-of-the-art. Monolingual models in combination with online Google machine translation models performed poorly compared to our proposed models as well as monolingual models, highlighting the problems that could be propagated with the use of machine translation models.

Recent work by Frank et al. (2018) on assessing multilingual multimodal image descriptions has found that descriptions generated in the target language are preferred over translations. In Table 5.4 we present the English image description ranking results of our best performing system against various state-of-the-art methods that use different techniques to encode both descriptions and images and map them in the joint space. It is also observed that visual features play a great role in learning better representations which is demonstrated in Table 5.4. There has been evidence when used extra visual information from ground truth bounding boxes or phrase annotations boosts the performance of image-ranking task especially for retrieving relevant descriptions when image is given as query.

For German, both PIVOT and PARALLEL with the asymmetric scoring function outperform monolingual models and symmetric similarity. We also observe that the German ranking experiments benefit the most from the multilingual signal. A reason for this could be that the German description corpus has many singleton words (more than 50% of the vocabulary, see Table 5.2) and English description mapping might have helped in learning better semantic embeddings or German which is morphologically richer than English and differs syntactically (e.g., in terms of word order).

These results suggest that the multilingual signal could be used to learn better multimodal embeddings, irrespective of the language. Our results also show that the asymmetric scoring function can help learn better embeddings. In Tables 5.6 and 5.7 we present top retrieved images for queries in English and German. As shown in the examples we found that all our models including the baseline model of Vendrov et al. (2016) are not efficient at interpreting numbers and quantifiers.

In Table 5.8 we present a few examples where PIVOT-ASYM and PARALLEL-ASYM models performed better on both the languages compared to baseline order embedding model even using descriptions of very different lengths as queries.

OE	
PIVOT-ASYM	
PARALLEL-ASYM	

Table 5.6: Two people riding a colorfully decorated bicycle

OE	
PIVOT-ASYM	
PARALLEL-ASYM	

Table 5.7: zwei junge Männer trommeln auf der Straße

Image	Descriptions	Image Rank		
		OE	PIVOT	PARALLEL
	2 Menschen auf der Straße mit Megafon	141	37	6
	two people in blue shirts are outside with a bullhorn	85	7	3
	ein Verkäufer mit weißem Hut und blauem Hemd , verkauft Kartoffeln oder ähnliches an Männer und Frauen	36	1	3
	at an outdoor market , a small group of people stoop to buy potatoes from a street vendor , who has his goods laid out on the ground	24	2	2

Table 5.8: The rank of the gold-standard image when using each German and English descriptions as a query on models trained using asymmetric similarity.

5.4.4 Word-query Retrieval

To understand whether our models are capable of retrieving images relevant to word based queries we queried using single words referring to objects (nouns), actions (verbs) and scenes. A few qualitative image retrieval results for queries referring to objects *dog, drums*, verbs *riding, playing* and scenes *night, beach* in both English and German are shown in Figure 5.5. We observe that our models are able to retrieve relevant images for all different word categories in both English and German. For example query *riding* retrieved images that represent vehicle riding whereas *reiten* one of the translations of verb riding into German retrieved images that represent animal riding meaning.

Prior work on neural word representation learning models by Mikolov et al. (2013b) has showed that learned word representations exhibit semantic regularities such as word analogies: $king - man + woman \; queen$. Similar regularities have been observed for multimodal representations learned for images and text (Kiros et al., 2014; Vendrov et al., 2016). We find that our multilingual multimodal representations using asymmetric similarity exhibit compositional regularity. In Figure 5.6 we analyse applicability of this regularity to combination languages and modalities using element-wise max (composition) operation for combination of English, German word queries.

5.4.5 Semantic Textual Similarity

Semantic Textual similarity (STS) determines the semantic equivalence between two texts. Given a pair of texts the task is to determine the semantic equivalence of the texts on a scale of $0-5$, higher scores indicates higher similarity between texts. This is a well studied task in natural language processing with its applications such as textual entailment, machine translation and question answering. To evaluate our model on the semantic textual similarity task (STS), we use the textual embeddings from our model to compute the similarity between a pair of sentences (image or video descriptions in this case).

We evaluate on video task from STS-2012 and image tasks from STS-2014, STS-2015 (Agirre et al. 2012, Agirre et al. 2014, Agirre et al. 2015). Each of these datasets are graded on a scale of $0-5$ based on how similar the two given sentences are to each other (annotation guidelines are shown in Figure 5.7). The video descriptions in the STS-2012 task are from the MSR video description corpus (Chen & Dolan, 2011) and the image descriptions in STS-2014 and 2015 are from UIUC PASCAL dataset (Rashtchian et al., 2010).

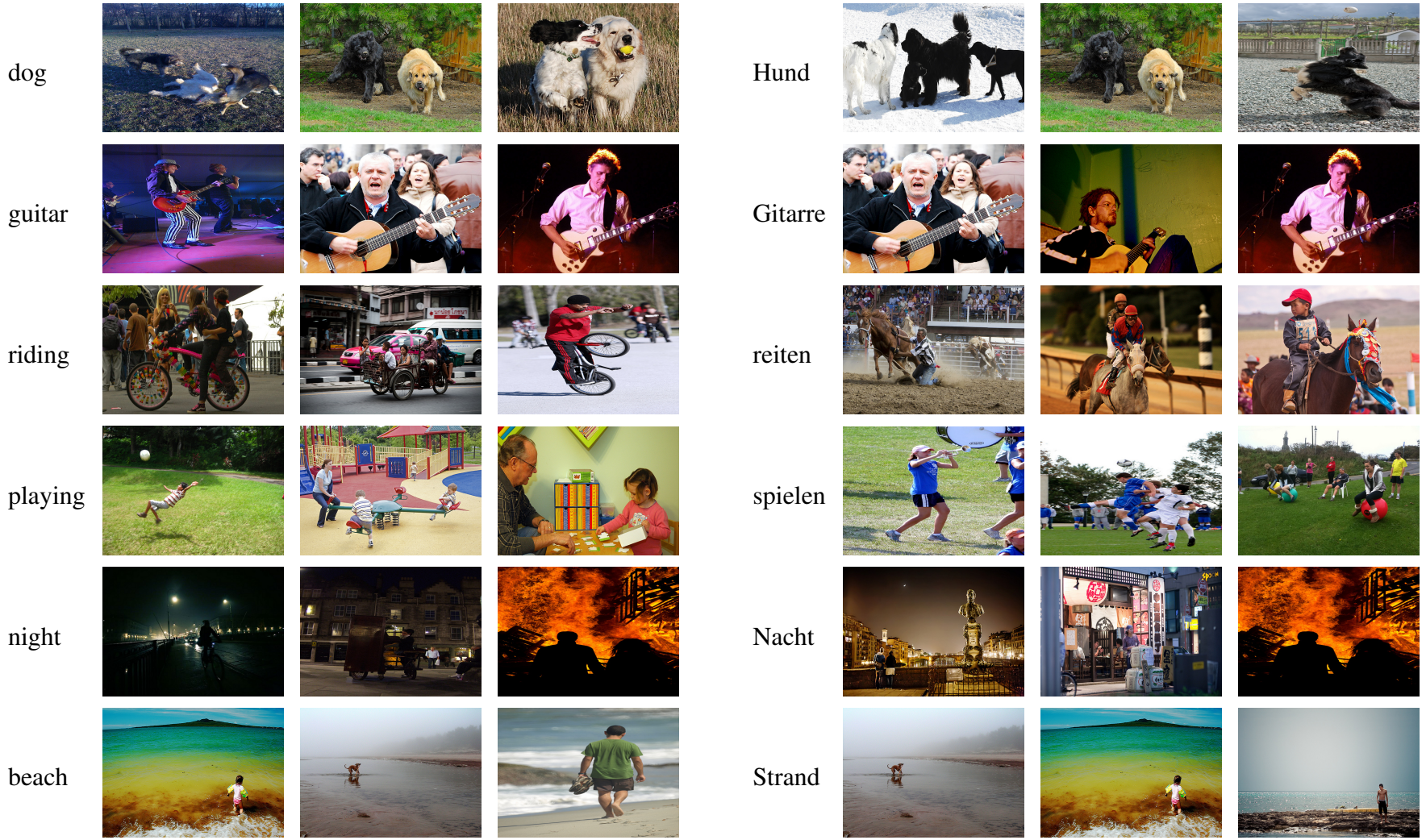


Figure 5.5: Top retrieved images from test using PARALLEL-ASYM model for both English and German word queries. We present image retrieval results for object: dog, drums, verb: riding, playing and scene: night, beach

Figure 5.6: Multilingual, multimodal compositional vector space arithmetic: regularities found with representations learned using PARALLEL-ASYM model for a combination of English and German word combination queries.

(5) **Completely equivalent**, as they mean the same thing

The bird is bathing in the sink.
Birdie is washing itself in the water basin.

(4) **Mostly equivalent**, but some unimportant details differ

Two boys on a couch are playing video games.
Two boys are playing a video game.

(3) **Roughly equivalent**, but some important information differs/missing

John said he is considered a witness but not a suspect.
"He is not a suspect anymore." John said

(2) **Not equivalent**, but share some details

They flew out of the nest in groups.
They flew into the nest together.

(1) **Not equivalent**, but are on the same topic

The woman is playing the violin.
The young lady enjoys listening to the guitar.

(0) **On different topics.**

The black dog is running through the snow.
A race car driver is driving his car through the mud.

Figure 5.7: Annotation guidelines to score how similar two given sentences are to each other from Agirre et al. (2013)

In Table 5.9, we present the Pearson correlation coefficients of our model predicted scores with the gold-standard similarity scores provided as part of the STS image/video description tasks. We compare with the best reported scores for the STS shared tasks, achieved by MLMME (Calixto et al., 2017b), paraphrastic sentence embeddings (Wieting et al., 2017), visual semantic embeddings (Kiros et al., 2014), and order embeddings (Vendrov et al., 2016). Wieting et al. (2017) uses neural machine translation to generate paraphrases for sentences via back-translation of bilingual sentence pairs. They propose Gated Recurrent Averaging Network (GRAN) that combines average embeddings of the word in the sentence with long short-term memory (LSTM) recurrent neural network to generate the representation of a sentence.

Model	VF	2012	2014	2015
Shared Task Baseline	–	29.9	51.3	60.4
STS Best System	–	**87.3**	83.4	86.4
GRAN (Wieting et al., 2017)	–	83.7	**84.5**	85.0
MLMME (Calixto et al., 2017b)	VGG19	–	72.7	79.7
VSE (Kiros et al., 2014)	VGG19	80.6	82.7	89.6
OE (Vendrov et al., 2016)	VGG19	82.2	84.1	90.8
PIVOT-SYM	VGG19	80.5	81.8	89.2
PARALLEL-SYM	VGG19	82.0	81.4	90.4
PIVOT-ASYM	VGG19	83.1	83.8	90.3
PARALLEL-ASYM	VGG19	84.6	**84.5**	**91.5**

Table 5.9: Results on Semantic Textual Similarity Image datasets (Pearson's $r \times 100$). Our systems that performed better than best reported shared task scores are in **bold**.

The shared task baseline is computed based on word overlap between the sentences and is high for both the 2014 and the 2015 dataset (51.3 and 60.4 respectively), indicating that there is substantial lexical overlap between the STS image description datasets. Our models outperform both the baseline system and the best system submitted to the shared task. For the 2012 video paraphrase corpus, our multilingual methods performed better than the monolingual methods showing that similarity across paraphrases can be learned using multilingual signals. Similarly, Wieting et al. (2017) have reported to learn better paraphrastic sentence embeddings with multilingual signals.

Overall, we observe that models learned using the asymmetric scoring function outperform the state-of-the-art on these datasets, suggesting that multilingual sharing is beneficial and our multilingual multimodal representation models can be used as off-the-shelf models to learn representations for sentences. Although the task has nothing to do with German, because our models can make use of datasets from different languages, we were able to train on significantly larger training dataset of approximately 145k descriptions. Calixto et al. (2017b) also train on a larger dataset like ours, but could not exploit this to their advantage. In Table 5.10 we present the example sentences with the highest and lowest difference between gold-standard and predicted semantic textual similarity scores using our best performing PARALLEL-ASYM model.

S1	S2	GT	Pred
Depressed woman sitting on couch	Older woman holding newborn baby	0.0	2.43
Black bird standing on concrete.	Blue bird standing on green grass.	1.0	4.2
The lamb is looking at the camera	A small bird standing on a log at the waters edge	0.0	1.95
Two zebras are playing.	Zebras are socializing.	4.2	1.2
Three goats are being rounded up by a dog.	Three goats are chased by a dog	4.6	4.5
Two green and white trains sitting on the tracks	Two green and white trains on tracks	4.4	4.62
A man is folding paper.	A woman is slicing a pepper.	0.6	0.6
A man sitting on a bench looking at a dog on a leash sitting on sidewalk	A woman holds a small baby while sitting on a sofa	0.0	1.2
Tan cows look closely at the camera	A white and grey cat in a bathroom sink looking at the camera	0.75	2.09

Table 5.10: Example sentences with gold-standard semantic textual similarity score and the predicted score using our best performing PARALLEL-ASYM model.

5.4.6 Crosslingual Image Description Task

The crosslingual image description task is proposed as part of the WMT multimodal machine translation task. An image is provided along with source language descriptions (English) and the task is to generate a description for the image in the target language (German). The descriptions part of Multi30k dataset was provided for training the models for this task.

We used out PIVOT and PARALLEL models to retrieve a German description for a given test image. We test this in two different scenarios (i) retrieve the top German description closest to the test image without using the English descriptions provided (ii) retrieve the top German description that is closest to both the image and the English descriptions provided. For a given test image, using our models (i) we encode the test image; (ii) we encode both the test image and the English descriptions provided. In the first scenario we extract the German description that is closest to the the test image from training set of German descriptions. In the second scenario, we also sort the top N retrieved German descriptions that are closer to the test image based on similarity to the English source descriptions. We pick the German descriptions which is closer to both the image and the English descriptions provided.

We present our model results in Table 5.11. We observed significant improvement by re-ranking the retrieved German descriptions that are closer to the provided English descriptions. For all our models we observed that using English descriptions for re-ranking improved almost 5 BLEU and METEOR points. The best METEOR score is observed for PIVOT-ASYM model, our best performing model for learning multilingual multimodal representations. However, all of our systems performed poorer than the multimodal RNN generation based system of Elliott et al. (2015) which was provided as a baseline for the crosslingual image description task.

5.4.7 Cross-lingual Retrieval

We evaluate our models on the cross-lingual retrieval task where a source language sentence is provided as a query and the system has to retrieve a set of relevant sentences in the target language which are closer to the source language sentence (closeness measured by similarlity metric, cosine: in cases of PARALLEL-SYM and PIVOT-SYM models and asymmetic similarity metric: in cases of PARALLEL-ASYM and PIVOT-ASYM). To evaluate our model this setup is ideal when the image is not present during test-time and the retrieval has to be performed solely based on the text queries. For

System	I	En	BLEU ↑	Meteor ↑	TER ↓
Grounded Translation (Elliott et al., 2015)	Y	N	**15.8**	**31.2**	76.4
UPC (Guasch & Costa-Jussà, 2016)	Y	Y	1.5	12.1	63.1
CUNI (Libovický et al., 2016)	Y	Y	1.2	13.1	73.3
PIVOT-SYM	Y	N	5.7	20.4	85.2
PARALLEL-SYM	Y	N	6.7	20.7	85.2
PIVOT-ASYM	Y	N	5.4	20.2	89.7
PARALLEL-ASYM	Y	N	6.1	20.0	92.0
PIVOT-SYM	Y	Y	10.3	25.6	74.9
PARALLEL-SYM	Y	Y	12.4	26.6	**72.5**
PIVOT-ASYM	Y	Y	11.3	27.0	74.3
PARALLEL-ASYM	Y	Y	11.4	27.8	74.9

Table 5.11: Results on WMT'16 Multimodal Machine Translation Task2; Column En denote whether English descriptions provided at test are used to re-rank the German descriptions or not.

this task we use the Translation corpus of Multi30k where there are 1000 parallel sentences in English an German. The image description corpus of Multi30k do not have alignments between English and German descriptions and cannot be used for our language retrieval experiments. We test this on both of our PARALLEL models PARALLEL-SYM and PARALLEL-ASYM.

To compare the accuracy of our PARALLEL-SYM and PARALLEL-ASYM models we present scores for Canonical Correlation Analysis and its variants. CCA based models have been used to map two views on related data into a shared embedding space which maximally correlates linear projections of both views. CCA has been shown to be very useful in closely related tasks such as learning bilingual text embeddings. Various CCA variants has been proposed to map two or more views into same space such as: Generalised Canonical Correlational Analysis (GCCA) which supports more than two views (Funaki & Nakayama, 2015a), Partial Canonical Correlational Analysis (PCCA): learns maximally correlated linear projections of two views conditioned on a shared third view (Rao, 1969), Deep Canonical Correlational Analysis (DCCA): a deep learning variant of CCA which learns non-linear projections of two views (Wang et al., 2015) and Deep Partial Canonical Correlational Analysis (DPCCA): a deep learning variant of PCCA which allows both conditioning on third view (images in our cases) and allows non-linear transformations on the data (Rotman et al., 2018).

	Recall@1		Recall@5	
Model	**EN→DE**	**DE→EN**	**EN→DE**	**DE→EN**
CCA (Hotelling, 1936)	76.40	70.40	91.60	88.50
PCCA (Rao, 1969)	78.50	73.70	**92.80**	90.40
GCCA (Funaki & Nakayama, 2015a)	69.90	69.00	87.20	87.90
DCCA (Wang et al., 2015)	61.90	62.10	82.80	82.50
DPCCA(Rotman et al., 2018)	**80.90**	**79.40**	**92.50**	91.20
PARALLEL-ASYM	77.20	76.30	**92.70**	**93.10**
PARALLEL-SYM	60.60	62.70	84.50	84.40

Table 5.12: Results for cross-lingual description retrieval on Multi30k Translations dataset; Column EN denote English descriptions and DE denote German.

We present Recall@1 and Recall@5 scores for EN-DE retrieval results and vice versa. Note that except for our model all the other models reported here are also trained on Multi30k translation corpus (29k parallel sentences between English and German and their images in few cases) whereas our models are trained on Image-English 145k mappings and Image-German 145k mappings. That is all the other models use parallel alignments with both languages whereas we rely on comparable data from an image description corpus. All of the models including ours is tested on the test corpus of the Multi30k translation data. In Table 5.12 we present scores for cross-lingual retrieval. We observe that DPCCA models achieve best scores for Recall@. However, our PARALLEL-ASYM model achieves best scores for Recall@5 on retrieval from both EN-DE and DE-EN.

We believe training on comparable data could be the reason for these scores. Since in comparable data we our models have never seen exact mappings or translation of English-German pairs but have been optimized to map related pairs closer to each other in the joint space. We also observe that similar to other ranking-experiments PARALLEL-ASYM model performed better than PARALLEL-SYM, indicating that asymmetric measure is more efficient than symmetric similarity measure . Despite training on comparable data our models perform competitively at Recall@1 compared to other state-of-the-art methods and achieve best score for the Recall@5 measure.

5.5 Conclusions

We proposed two new models that jointly learn multilingual multimodal representations using the image as a pivot between languages. We introduced new objective functions that can exploit similarities between images and descriptions across languages. We obtained state-of-the-art results on two tasks: image-description ranking and semantic textual similarity and competitive results on crosslingual image description generation and crosslingual retrieval tasks. Our image-description ranking experiments show that multi-task or joint learning is a potential direction to explore to extend multimodal applications in multiple languages.

Kádár et al. (2018) have extened our models to more than two languages and presented experiments of using translations (parallel-data) vs. comparable data for learning multulingual multimodal representations. Their experiments suggest that multilingual multimodal representations can be trained equally well on either translations or comparable sentence pairs. Their results on more than two languages suggest that annotating the same set of images in multiple language enables further improvements. However, the question of which languages when jointly trained benefits the most is still to be addressed and we envision this would be a possible future direction of this work to explore.

Our results on semantic textual similarity suggest that multilingual multimodal representations learned are indeed useful for natural language understanding tasks. Our experiments on crosslingual image description shows the extent to which our multilingual multimodal representations capture semantic relatedness of image and text in multiple languages. Similarly our cross lingual experiments suggest that our models trained on comparable image descriptions perform competitvely with models trained on parallel sentences for the task of cross-lingual image retrieval.

Overall, we observe that exploiting multilingual and multimodal resources can help in learning better semantic representations that are useful for various multimodal natural language understanding tasks. In the future, we would like to explore how this framework can benefit other tasks such as bridging language, speech and vision.

Chapter 6

Conclusions and Future Directions

This thesis has demonstrated the benefits of utilizing visual context for both action recognition and multilingual multimodal representation learning. In Chapter 2 we summarized the use of visual context from images, language context from image descriptions for identifying verbs that denote actions in images and linguistic resources such as OntoNotes to distinguish different meanings of verbs depicted in the images.

In Chapter 3 we show that salient regions identified my convolutional neural network models to identify visual verbs or actions correlate with regions fixated by humans while performing an action classification task. In Chapter 4 we demonstrate the usefulness of visual information to resolve lexical ambiguity across languages. Additionally, we also show that visual disambiguation can be used to improve the performance of a machine translation system on image descriptions. Finally, in Chapter 5 we propose models to utilize images as a pivot between languages to utilize resources form other languages to ground and learn representations for cross-lingual search and image search. In this chapter, we address some of the limitations of our work and we conclude with a broader discussion of promising topics of future research in this area.

6.1 Limitations

One of the limitations observed in our sense disambiguation task is scaling to a larger number of verbs. There are three main challenges involved in scaling to a larger number of verbs: (i) collecting annotations especially sense annotations require expert annotators; (ii) reporting bias issues or long-tail distribution of images found on the web; (i) Not all visual actions are visualisable in images and some actions might require videos to understand and disambiguate the meaning

Both in our VerSe and MultiSense collection we observed that identifying the visual senses of the verb in a language or across languages is a difficult task. Our inter-annotator agreement scores for these tasks show lower agreement compared to annotating images with noun senses (Deng et al., 2009). A reason for this could be a large number of senses for verbs, high variability in the image and the meaning of the verb depending on the scene and objects involved. In Chapter 4, we observed that inter-annotator agreement for translation is lower compared to sense tagging in Chapter 2. This suggests that sense annotation for verbs is not easy to crowdsource on large scale and requires expert annotators.

In this thesis, we have mainly explored visual context from the image whereas many actions require more than a single frame i.e., video to understand the semantics. For example verbs such as *run* (motion verb), *travel* (location change verb), spill (state change verb) are better understood with video signal. We believe working with videos would cover larger set of verbs. However, working with videos might involve greater annotation efforts.

One major limitation of our multilingual multimodal representation learning model is we train our models on a relatively smaller dataset. Other parallel works which learn joint embedding spaces of images and text have used MSCOCO image-description data which is much larger than the Flickr30k dataset. In future, we would like to utilize larger image description datasets (for example MSCOCO) either to pre-train our models or use this as additional training data. Another limitation is our models use off-the-shelf image representations from pre-trained models on ImageNet. We believe end-to-end training using images would have helped in learning better representations. Similar ideas have been explored in image description generation (Lu et al., 2017).

6.2 Future Directions

6.2.1 Extensions of VSD

In this thesis, we explored visual sense disambiguation as a standalone task. An obvious application for this would be image search: recall Figure 2.2, which depicts a search result obtained with the verb *sit* as query. If the search engine had access to verb sense disambiguation for images, then it would be able to cluster the search results based on verb senses, rather than forming groups based on image or query similarity.

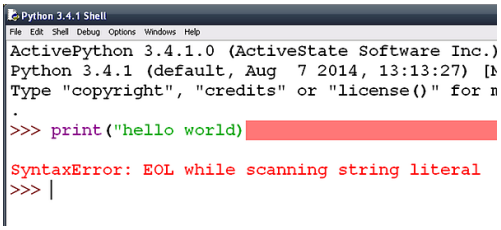
Other language and vision task that is likely to benefit include image description and visual question answering. An image description system that has access to verb prediction and sense disambiguation can make sure that it outputs only descriptions that are compatible with the verb senses that are attested in the image it tries to describe. A simple re-ranking architecture could be used to implement this: We take an existing image description system, use it to generate a set of candidate descriptions for a given image, and then re-rank the descriptions based on the output of our verb prediction and VSD models. In a similar fashion, VSD could be used to re-rank the output of a visual question answering system (or the VSD scores could simply serve as a feature).

We could apply our sense disambiguation models to zero-shot or one-shot learning, i.e. the classification of actions for which there are very few training samples, as low as one example, or no training examples at all. Since a large set of actions seem to have a long-tailed distribution over the images in the action recognition datasets or over the images found on the web.

Furthermore, as discussed in this thesis, lexical databases such as OntoNotes, FrameNet provide information about the participants and role of each participant in actions. Most of the existing action recognition datasets do not utilize this information except recent work by Yatskar et al. (2016). A straightforward extension would be to incorporate verbs with multiple visual senses in the Situation Recognition task (Yatskar et al., 2016) that not only generalizes to situations involving different objects and scenes but also different meanings of the verb (e.g., include polysemous verbs such as *play*).

In Chapter 4 we have shown that visual information could be used to identify the translation of a verb from the source language to the target language, i.e. resolving lexical ambiguity across languages. We also demonstrated that this information could be plugged into a neural machine translation model and build better translation systems for image descriptions. An interesting study would be to extend this to translation of other part-of-speech categories such as nouns and adjectives.

A bright yellow honey comb with a bee making honey. It's hexagonal in shape.
It -> honeycomb (coreference)



Are you look at the bug on the screen? It looks like tiny one.
bug -> problem that needs fixing (lexical ambiguity)

Figure 6.1: Examples depicting how visual information could be helpful in resolving coreference and lexical ambiguity in multi-sentence machine translation.

A recent work by Lala & Specia (2018) introduced multimodal lexical translation task and dataset which covers ambiguous words from many part-of-speech categories. Another extension and useful application of cross lingual sense disambiguation is to model coreference and context in discourse based machine translation where you are translating more than a single sentence (Bawden et al., 2018). Apart from linguistic context visual context could be used to generate correct translations or gender agreements between sentences. In Figure 6.1 we show a couple of examples where visual information could be helpful in resolving coreference and lexical ambiguity in discourse machine translation problem.

6.2.2 Visual Context for Common Sense Learning

Another extension would be to study and model affordance. Affordance is a well-defined concept in psychology, is considered as a relation between an object/environment and an organism that affords the opportunity for that organism to perform an action (Gibson, 1977), a critical component of common sense. Recently Chao et al. (2015c) proposed a model to address semantic affordance problem, given an object determining whether an action can be performed by a human on it. This problem can also be viewed as a sub-problem of common sense learning where we can identify or learn the possible set of meaningful or plausible interactions that could happen between

objects.

For example, a human can perform the action of "carry :: transport, move while supporting" on objects like "bag" and "dog" but not with objects like "elephant". Whereas a human can carry a toy elephant. Such things seem obvious to humans whereas it is difficult to automate a system to learn such information. One reason for this humans rarely state the obvious things such as adult elephants are larger and heavier than humans whereas most dogs and bags are smaller than human. This is a well studied phenomena known as reporting bias Van Durme (2009). This information could be very helpful in robot automation since semantic affordance varies depending on various attributes such as size, texture, weight and multimodal information could be helpful in determining it.

A preliminary work on this by Forbes & Choi (2017) have shown encouraging results on extracting such common sense knowledge just from text. To the best of our knowledge, there is no existing knowledge-base which explicitly encodes such information. We believe, we can learn such knowledge from multimodal resources using a combination of a large body of text or existing text-based semantic-knowledge resources such as WordNet, FrameNet and visual knowledge bases such as ImageNet or images from the web.

In this work, we have only studied actions and verbs that denote actions in images. However, a large group of motion-based actions require more than a single frame to understand the actions such as differentiating between *opening the door* and *closing the door*, *running* and *jogging*. Also, humans have the innate ability do to multi-tasking, i.e. performing multiple actions at the same time. For example, talking to another person while searching for a file in the cabinet and drinking the coffee, all at the same time. For the systems to understand and process such combinations of complex actions it is necessary to process videos and to focus on recognizing multiple actions and relationships among them at a time. Also, a large set of actions are intricately connected i.e., there is a causal relationship between a pair of sequential actions. For example, a ball that is thrown into the air will fall on the ground or when you bite an apple you chew it and then swallow it.

We believe extending verb identification and distinguishing between fine-grained analysis of verbs to videos will provide greater insights into video understanding and take a step further towards common sense learning. In Chapter 3 we have shown that there is a high correlation between salient features predicted by verb prediction models and human observers. An extension would be to see if this extends to videos and whether this information could be used to identify future actions in an event or

identifying the trajectory of an object.

6.2.3 Multilingual Multimodal Representation Learning

We have shown that we could learn better multilingual multimodal representations by using an image as a bridge between image descriptions in more than one language. A straightforward extension would be to extend this work to languages with limited resources in multimodal space i.e., which do not have large number of image and sentence pairs. Similar ideas have been shown to be effective in part-of-speech tagging and machine translation where significant improvement was seen when a low-resource language is jointly trained with high-resource language. In this thesis, we have shown that models built using comparable data are capable of contributing to better representations. However, if there exists parallel data, i.e. translations between sentences these representations could be generalized to many other language tasks such as crosslingual search, multilingual conversational systems which enable querying or conversation in multiple languages at the same time. A recent work by Kádár et al. (2018) have shown how this models could be extended to more than two languages as well as an analysis of using translations vs. comparable data for learning multulingual multimodal representations. Their analysis also include ablation studies how our proposed approach is effective in case of low-resource settings. Kádár et al. (2018) have used more advanced visual features and their results show much more consistency using improvements in multilingual multimodal representations when compared to single language multimodal representations. This is encouraging showing that stronger visual signal results in better multimodal representations.

We have used words as units in our recurrent neural network in both out models for learning representations. If the data is sparse, this is not an efficient method since many words might be singletons or occurring very few times. One extension of our word could be to use sub-word units instead of words as units in encoding text. Also, our models use separate encoders for each language, if we learn joint sub-word units we could use a single recurrent neural network encoder to encode the text. A recent study on learning universal representations which uses a single encoder to encode many languages have shown boost in the performance for machine translation task (Johnson et al., 2017; Schwenk, 2018).

To compare our models with previous existing works on single language multimodal representations (See Table 5.4) we did not experiment with various visual features.

Another extension would be to test the contribution of better visual features in our models experimenting with ResNet or InceptionNet features which perform better than our VGG features on image classification and other tasks. An additional advantage is both ResNet and InceptionNet has fewer dimensions than VGG, i.e. our models have to learn fewer number of parameters.

In this thesis, we have only scratched the surface of multimodal applications which could potentially benefit from extending it to multiple languages. This could be easily extended to multilingual or multiple language image description (some preliminary work in this space was explored by Elliott et al. (2015)), multilingual visual question answering, multilingual visual dialog etc. We believe there is a wide range of possibilities and extensions for representation learning with vision as a bridge between languages or language and speech. Following our work presented in Chapter 5 (published as Gella et al. (2017a)). Harwath et al. (2018) have shown that image pivoting is useful to learn representations for speech in multiple languages. Their results are encouraging and open many possibilities of learning joint representations for speech, text and vision. Our work and the recent extensions to it by Kádár et al. (2018) only handled languages that are similar i.e., English, German and French whereas work by Harwath et al. (2018) have not only shown that our proposed models works for novel modalities like speech but also works on languages that are very different from each other. Their multimodal audio-visual retrieval is studied on English and Hindi which have completely different grammar and word order. This opens up many interesting use cases for multimodal multilingual research.

Appendix A

A.1 Visual Sense Visualness Annotations

Our visualness annotation of OntoNotes lists 921 senses for our 148 target verbs. Out of which our annotators marked 504 senses are depictable pr visual. Below, we present sense definitions and visual binary label for each of the sense for verbs *serve* and *play*.

lemma	sense num	definition	ontonotes sense examples	label
serve	1		This table serves both as a desk and a work bench. . That sewage plant serves the entire coastal community.	False
serve	3	perform a duty	He served in Congress for two terms. They served their country nobly.	False
serve	4	dish out, hand out something, often food	They are serving mint juleps on the veranda. Her club serves meals to the homeless on Thursdays.	True
serve	5	spend time in prison	He served twenty years in prison for armed robbery. The prisoners were released before they had served their full terms.	True
serve	6	mate with, sexual reproduction	This pedigree stallion served three brood mares.	True
serve	7	put a ball into play	It was Mary's turn to serve. That tennis star has been serving erratically lately.	True

play	1	engage in a fun or recreational (child-like) activity	The children are playing across the street. Life is short, play hard. Let's play hide-and-seek.	True
play	2	engage in or make moves related to competition or sport	They played cards far into the night. Do you want to play tennis with me tomorrow? Princeton plays Yale this weekend.	True
play	3	behave in a certain way; have a specific effect or outcome	Money is playing a big role in his decision to take the job. That boy played no part in the vandalism. I think we should play it safe.	False
play	4	perform or transmit music	The band played all night long. The radio is playing my favorite song. She played some very difficult Beethoven at the recital.	True
play	5	perform/act a role, pretend	He usually plays a villain in films. She plays the lead in Evita. The show only played three nights before closing.	True
play	6	wager, bet	He used to play the ponies a lot. They played the casinos every night in Las Vegas. I'd play my money on the horse from Tennessee.	False
play	7	toy, fiddle, or trifle with; act without the expected seriousness	He plays the stockmarket a little on the side. She really knows how to play on their emotions, doesn't she? She has played with the idea of starting a dating service.	False
play	8	be interpreted or received	That campaign speech won't play well in Peoria.	False

play	9	move freely (usually within a bounded space)	I think this steering wheel is playing too much. The city lights played over the still waters. The spotlights played on the politicians.	False
play	10	run or operate	The fountains played all day.	False
play	11	FISHING-exhaust by allowing to pull on the line	John knows how to play a hooked fish.	True
play	12.1	PLAY ALONG-cooperate or pretend to cooperate	He decided to play along with the burglars for the moment. I don't know where he's going with it, but just play along for now.	False
play	12.2	PLAY ALONG-musically accompany	Children love to play along on the piano and sing their favorite songs. Is it ok to play the song along with the tape?	False
play	12.3	PLAY AROUND-work or deal with in an amateurish or casual manner	He plays around with investments but he never makes any money. I've been playing around with the idea of writing poetry for a while now. Will you stop playing around? We are trying to get some work done here.	False
play	12.4	PLAY AROUND-commit adultery	He says that he is not merely interested in playing around with her. He plays around a lot.	False
play	12.5	PLAY BACK-reproduce a recording as on a recorder	They played back the conversation to show that their client was innocent. Play back the tape you just recorded and listen carefully. She played back the incident over and over again in her head.	False

play	12.6	PLAY DOWN-understate importance or quality of	A lawyer by profession, he knew he must mentally play down the danger. Helen played down her disability despite its devastating effect. She played down her influence on domestic politics.	False
play	12.7	PLAY OFF-set into opposition or rivalry	The winners will play off against each other in the Championship Cup. Hungarian minimalism plays off against Polish expressionism. Watt and the guitarist played off of each other putting on a great performance!	False
play	12.8	PLAY OUT-happen or develop; go from beginning to end	I wonder how this debate will play out. The way the conference played out last year, you have to be ready every weekend.	False
play	12.9	PLAY UP-emphasize, highlight, or exaggerate	No need to play the story up. The magnitude of the event is implicit in the facts. James Plum played up his life's setbacks as if they were 'gifts...even treasures'.	False
play	12.10	PLAY UP-meet a standard or expectation	Gillian Apps played up to her title, as she propelled her team to victory. He's had some weeks where he hasn't played up to his expectations or ours.	False
play	12.11	PLAY UP-ingratiate oneself, often with insincere behavior	It's unbelievable how she plays up to her supervisors. The cast members so obviously played up to the camera.	False
play	12.12	idiomatic expressions		False

A.2 VerSe Annotations

Below, we present images in our Verse dataset grouped according to same sense labels for the verbs *play* and *serve*. For the verb *play*, we present images grouped according to 3 senses namely: *engage in a fun or recreational activity*, *engage in or make moves related to competition or sport*, *performing music*. For the verb *serve*, we present images grouped according to senses namely: *serving food* and *serving ball*.

play#1: engage in a fun or recreational (childlike) activity

play#2: engage in or make moves related to competition or sport

play#4: perform or transmit music

serve#4: dishout, hand out something often food

serve#7: put a ball into play

A.3 Verb Localizations

Additional to examples presented in Chapter 2 here we present visualisations of different images for the verbs *fly*, *smile* and *feed*. Despite involving various type of objects, number of objects and different type of scenes our model predicted and localized the most relevant region of the image.

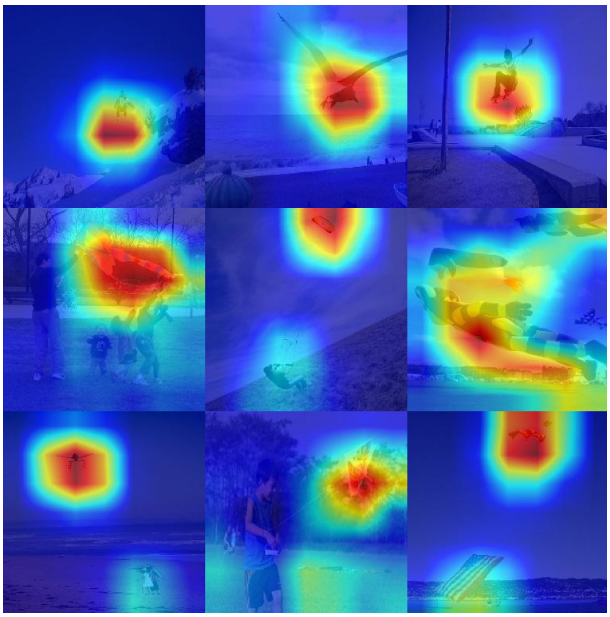


Figure A.1: Localizations for predicted verb *fly*

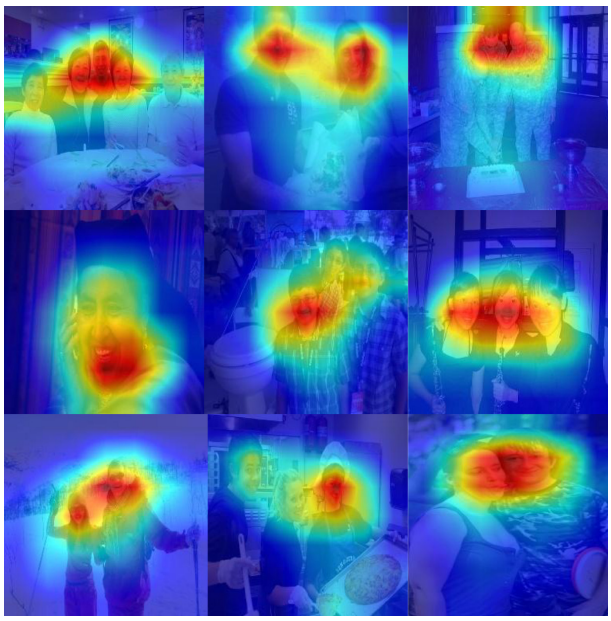


Figure A.2: Localizations for predicted verb *smile*

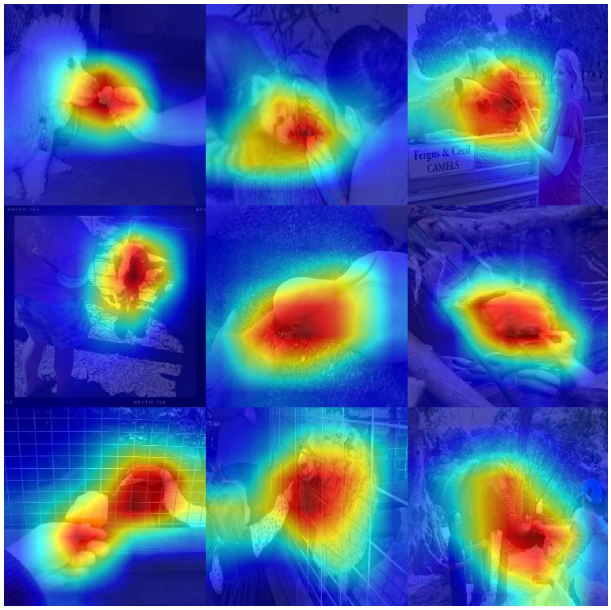


Figure A.3: Localizations for predicted verb *feed*

Bibliography

Agirre, Eneko, Diab, Mona, Cer, Daniel, and Gonzalez-Agirre, Aitor. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 385–393. Association for Computational Linguistics, 2012.

Agirre, Eneko, Cer, Daniel, Diab, Mona, Gonzalez-Agirre, Aitor, and Guo, Weiwei. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pp. 32–43, 2013.

Agirre, Eneko, Banea, Carmen, Cardie, Claire, Cer, Daniel, Diab, Mona, Gonzalez-Agirre, Aitor, Guo, Weiwei, Mihalcea, Rada, Rigau, German, and Wiebe, Janyce. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 81–91. Association for Computational Linguistics, 2014.

Agirre, Eneko, Baneab, Carmen, Cardiec, Claire, Cerd, Daniel, Diabe, Mona, Gonzalez-Agirrea, Aitor, Guof, Weiwei, Lopez-Gazpioa, Inigo, Maritxalara, Montse, Mihalceab, Rada, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 252–263, 2015.

Andrew, Galen, Arora, Raman, Bilmes, Jeff A., and Livescu, Karen. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 1247–1255, 2013a.

Andrew, Galen, Arora, Raman, Bilmes, Jeff A., and Livescu, Karen. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 1247–1255, 2013b.

Andriluka, Mykhaylo, Pishchulin, Leonid, Gehler, Peter, and Schiele, Bernt. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. VQA: visual question answering. In *2015*

IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 2425–2433, 2015.

Baker, Collin F, Fillmore, Charles J, and Lowe, John B. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pp. 86–90. Association for Computational Linguistics, 1998.

Barnard, Kobus and Johnson, Matthew. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1-2):13–30, 2005.

Barnard, Kobus, Johnson, Matthew, and Forsyth, David. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data-Volume 6*, pp. 1–5. Association for Computational Linguistics, 2003.

Barrett, Maria and Søgaard, Anders. Using reading behavior to predict grammatical functions. In *EMNLP Workshop on Cognitive Aspects of Computational Language Learning*, Lisbon, Portugal, 2015.

Barrett, Maria, Bingel, Joachim, Keller, Frank, and Søgaard, Anders. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*, pp. 579–584, 2016a.

Barrett, Maria, Keller, Frank, and Søgaard, Anders. Cross-lingual transfer of correlations between parts of speech and gaze features. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 1330–1339, 2016b.

Barrett, Maria, Bingel, Joachim, Hollenstein, Nora, Rei, Marek, and Søgaard, Anders. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 302–312, 2018.

Bawden, Rachel, Sennrich, Rico, Birch, Alexandra, and Haddow, Barry. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pp. 1304–1313, 2018.

Bergsma, Shane and Van Durme, Benjamin. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, pp. 1764. Citeseer, 2011.

Bernardi, Raffaella, Cakici, Ruket, Elliott, Desmond, Erdem, Aykut, Erdem, Erkut, Ikizler-Cinbis, Nazli, Keller, Frank, Muscat, Adrian, and Plank, Barbara. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artifical Intelligence Research*, 55:409–442, 2016.

Brody, Samuel and Lapata, Mirella. Good neighbors make good senses: Exploiting distributional similarity for unsupervised wsd. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 65–72. Association for Computational Linguistics, 2008.

Bruni, Elia, Tran, Nam-Khanh, and Baroni, Marco. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

Bylinskii, Zoya, Judd, Tilke, Borji, Ali, Itti, Laurent, Durand, Frédo, Oliva, Aude, and Torralba, Antonio. Mit saliency benchmark. http://saliency.mit.edu/, 2016.

Caglayan, Ozan, Aransa, Walid, Bardet, Adrien, García-Martínez, Mercedes, Bougares, Fethi, Barrault, Loïc, Masana, Marc, Herranz, Luis, and van de Weijer, Joost. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation*, pp. 432–439, Copenhagen, Denmark, September 2017.

Caglayan, Ozan, Bardet, Adrien, Bougares, Fethi, Barrault, Loïc, Wang, Kai, Masana, Marc, Herranz, Luis, and van de Weijer, Joost. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 597–602, 2018.

Calixto, Iacer, Liu, Qun, and Campbell, Nick. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1913–1924, 2017a.

Calixto, Iacer, Liu, Qun, and Campbell, Nick. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*, 2017b.

Carpuat, Marine and Wu, Dekai. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, volume 7, pp. 61–72, 2007.

Chandar, Sarath, Khapra, Mitesh M., Larochelle, Hugo, and Ravindran, Balaraman. Correlational neural networks. *Neural Computation*, 28(2):257–285, 2016.

Chao, Yu-Wei, Wang, Zhan, He, Yugeng, Wang, Jiaxuan, and Deng, Jia. HICO: A benchmark for recognizing human-object interactions in images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1017–1025, 2015a.

Chao, Yu-Wei, Wang, Zhan, Mihalcea, Rada, and Deng, Jia. Mining semantic affordances of visual object categories. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4259–4267, 2015b.

Chao, Yu-Wei, Wang, Zhan, Mihalcea, Rada, and Deng, Jia. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4259–4267, 2015c.

Chen, David L and Dolan, William B. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 190–200. Association for Computational Linguistics, 2011.

Chen, Xinlei, Fang, Hao, Lin, Tsung-Yi, Vedantam, Ramakrishna, Gupta, Saurabh, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015a.

Chen, Xinlei, Ritter, Alan, Gupta, Abhinav, and Mitchell, Tom M. Sense discovery via co-clustering on images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 5298–5306, 2015b.

Clarke, Alasdair DF and Tatler, Benjamin W. Deriving an appropriate baseline for describing fixation behaviour. *Vision research*, 102:41–51, 2014.

Cornia, Marcella, Baraldi, Lorenzo, Serra, Giuseppe, and Cucchiara, Rita. A deep multi-level network for saliency prediction. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 3488–3493. IEEE, 2016.

Dalmaijer, Edwin S, Mathôt, Sebastiaan, and Van der Stigchel, Stefan. Pygaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods*, 46(4):913–921, 2014.

Das, Abhishek, Agrawal, Harsh, Zitnick, Larry, Parikh, Devi, and Batra, Dhruv. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 932–937, 2016.

Delaitre, Vincent, Laptev, Ivan, and Sivic, Josef. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*, 2010.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Li, Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255, 2009.

Denkowski, Michael and Lavie, Alon. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, June 2014.

Donahue, Jeffrey, Anne Hendricks, Lisa, Guadarrama, Sergio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

Dorr, Michael and Vig, Eleonora. Saliency prediction for action recognition. In *Visual Content Indexing and Retrieval with Psycho-Visual Models*, pp. 103–124. Springer, 2017.

Elliott, Desmond and Kádár, Ákos. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pp. 130–141, 2017.

Elliott, Desmond, Frank, Stella, and Hasler, Eva. Multi-language image description with neural sequence models. *arXiv preprint arXiv:1510.04709*, 2015.

Elliott, Desmond, Frank, Stella, Sima'an, Khalil, and Specia, Lucia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016*, 2016.

Elliott, Desmond, Frank, Stella, Barrault, Loïc, Bougares, Fethi, and Specia, Lucia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, WMT 2017*, pp. 215–233, 2017.

Everingham, Mark, Gool, Luc J. Van, Williams, Christopher K. I., Winn, John M., and Zisserman, Andrew. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Everingham, Mark, Eslami, S. M. Ali, Gool, Luc Van, Williams, Christopher K. I., Winn, John M., and Zisserman, Andrew. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest, Srivastava, Rupesh K, Deng, Li, Dollár, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, Margaret, Platt, John C, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482, 2015a.

Fang, Hao, Gupta, Saurabh, Iandola, Forrest N., Srivastava, Rupesh K., Deng, Li, Dollár, Piotr, Gao, Jianfeng, He, Xiaodong, Mitchell, Margaret, Platt, John C., Zitnick, C. Lawrence, and Zweig, Geoffrey. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1473–1482, 2015b.

Firat, Orhan, Sankaran, Baskaran, Al-Onaizan, Yaser, Yarman-Vural, Fatos T., and Cho, Kyunghyun. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 268–277, 2016.

Forbes, Maxwell and Choi, Yejin. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 266–276, 2017.

Frank, Stella, Elliott, Desmond, and Specia, Lucia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413, 2018.

Frome, Andrea, Corrado, Greg S, Shlens, Jon, Bengio, Samy, Dean, Jeff, Mikolov, Tomas, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Funaki, Ruka and Nakayama, Hideki. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 585–590, 2015a.

Funaki, Ruka and Nakayama, Hideki. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal*, 2015b.

Ge, Gary, Yun, Kiwon, Samaras, Dimitris, and Zelinsky, Gregory J. Action classification in still images using human eye movements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–23, 2015.

Gella, Spandana and Keller, Frank. An analysis of action recognition datasets for language and vision tasks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 2: Short Papers*, pp. 64–71, 2017.

Gella, Spandana, Lapata, Mirella, and Keller, Frank. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 182–192, 2016.

Gella, Spandana, Sennrich, Rico, Keller, Frank, and Lapata, Mirella. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Short Papers*, pp. 2829–2835, Copenhagen, 2017a.

Gella, Spandana, Sennrich, Rico, Keller, Frank, and Lapata, Mirella. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pp. 2829–2835, 2017b.

Gibson, James J. The theory of affordances. *In Perceiving, Acting, and Knowing*, 1977.

Girshick, Ross B., Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 580–587, 2014.

Gong, Yunchao, Wang, Liwei, Hodosh, Micah, Hockenmaier, Julia, and Lazebnik, Svetlana. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pp. 529–545. Springer, 2014a.

Gong, Yunchao, Wang, Liwei, Hodosh, Micah, Hockenmaier, Julia, and Lazebnik, Svetlana. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 529–545, 2014b.

Grubinger, Michael, Clough, Paul, Müller, Henning, and Deselaers, Thomas. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, volume 5, pp. 10, 2006.

Guasch, Sergio Rodríguez and Costa-Jussà, Marta R. Wmt 2016 multimodal translation system description based on bidirectional recurrent neural networks with double-embeddings. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pp. 655–659, 2016.

Gupta, Abhinav, Kembhavi, Aniruddha, and Davis, Larry S. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

Gupta, Saurabh and Malik, Jitendra. Visual semantic role labeling. *CoRR*, abs/1505.04474, 2015.

Hahn, Michael and Keller, Frank. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 85–95, 2016.

Hardoon, David R., Szedmák, Sándor, and Shawe-Taylor, John. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Hartmann, Mareike and Søgaard, Anders. Limitations of cross-lingual learning from image search. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pp. 159–163, 2018.

Harwath, David, Chuang, Galen, and Glass, James. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4969–4973. IEEE, 2018.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, Ruidan, Lee, Wee Sun, Ng, Hwee Tou, and Dahlmeier, Daniel. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 388–397, 2017.

Helcl, Jindřich, Libovickỳ, Jindřich, and Varis, Dusan. Cuni system for the wmt18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 616–623, 2018.

Helcl, Jindřich and Libovický, Jindřich. Cuni system for the wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 450–457, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Henderson, John M. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003.

Hewitt, John, Ippolito, Daphne, Kriz, Brendan Callahan Reno, and Callison-Burch, Derry Wijaya Chris. Learning translations via images with a massively multilingual image dataset. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 345–356. Association for Computational Linguistics, 2018.

Hieber, Felix, Domhan, Tobias, Denkowski, Michael, Vilar, David, Sokolov, Artem, Clifton, Ann, and Post, Matt. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*, December 2017. URL https://arxiv.org/abs/1712.05690.

Hitschler, Julian, Schamoni, Shigehiko, and Riezler, Stefan. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 4188–4192, 2015.

Hodosh, Micah A. *Natural language image description: data, models, and evaluation*. University of Illinois at Urbana-Champaign, 2015.

Hokamp, Chris and Liu, Qun. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1535–1546, 2017.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Hovy, Eduard H., Marcus, Mitchell P., Palmer, Martha, Ramshaw, Lance A., and Weischedel, Ralph M. Ontonotes: The 90% solution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*, pp. 57–60, 2006.

Huang, Po-Yao, Liu, Frederick, Shiang, Sz-Rung, Oh, Jean, and Dyer, Chris. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016*, pp. 639–645, 2016.

Ikizler, Nazli, Cinbis, Ramazan Gokberk, Pehlivan, Selen, and Duygulu, Pinar. Recognizing actions from still images. In *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pp. 1–4, 2008.

Ikizler-Cinbis, Nazli and Sclaroff, Stan. Object, scene and actions: Combining multiple features for human action recognition. In *European conference on computer vision*, pp. 494–507. Springer, 2010.

Jaimes, Alejandro, Pelz, Jeff B., Grabowski, Tim, Babcock, Jason S., and Chang, Shih-Fu. Using human observer eye movements in automatic image classifiers. In *Human Vision and Electronic Imaging VI, San Jose, CA, USA, January 20, 2001*, pp. 373–384, 2001.

Jauhar, Sujay Kumar, Dyer, Chris, and Hovy, Eduard H. Ontologically grounded multi-sense representation learning for semantic vector space models. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 683–693, 2015.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross B., Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pp. 675–678, 2014.

Jiang, Ming, Huang, Shengsheng, Duan, Juanyong, and Zhao, Qi. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1072–1080, 2015.

Johnson, Melvin, Schuster, Mike, Le, Quoc V., Krikun, Maxim, Wu, Yonghui, Chen, Zhifeng, Thorat, Nikhil, Viégas, Fernanda B., Wattenberg, Martin, Corrado, Greg, Hughes, Macduff, and Dean, Jeffrey. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 5:339–351, 2017.

Kádár, Ákos, Elliott, Desmond, Côté, Marc-Alexandre, Chrupala, Grzegorz, and Alishahi, Afra. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pp. 402–412, 2018.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137, 2015a.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137, 2015b.

Karpathy, Andrej, Joulin, Armand, and Li, Fei Fei F. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pp. 1889–1897, 2014.

Karthikeyan, S, Jagadeesh, Vignesh, Shenoy, Renuka, Ecksteinz, Miguel, and Manjunath, BS. From where and how to what we see. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 625–632, 2013.

Kiela, Douwe and Bottou, Léon. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 36–45, 2014.

Kiela, Douwe, Hill, Felix, Korhonen, Anna, and Clark, Stephen. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *ACL (2)*, pp. 835–841, 2014.

Kiela, Douwe, Vulic, Ivan, and Clark, Stephen. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. ACL, 2015.

Kilgarrif, Adam. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proc. of the first international conference on language resources and evaluation*, pp. 581–588, 1998.

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

Klein, Benjamin, Lev, Guy, Sadeh, Gil, and Wolf, Lior. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014.

Klerke, Sigrid, Goldberg, Yoav, and Søgaard, Anders. Improving sentence compression by learning to predict gaze. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, CA, 2016.

Kruthiventi, Srinivas S. S., Gudisa, Vennela, Dholakiya, Jaley H., and Babu, R. Venkatesh. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5781–5790, 2016.

Kümmerer, M, Theis, L, and Bethge, M. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *International Conference on Learning Representations (ICLR 2015)*, pp. 1–12, 2014.

Lala, Chirag and Specia, Lucia. Multimodal lexical translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*, may 2018.

Lazaridou, Angeliki, Bruni, Elia, and Baroni, Marco. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1403–1414, 2014.

Lazaridou, Angeliki, Pham, Nghia The, and Baroni, Marco. Combining language and vision with a multimodal skip-gram model. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 153–163, 2015.

Le, Dieu Thu, Bernardi, Raffaella, and Uijlings, Jasper. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp. 231–238. ACM, 2013a.

Le, Dieu-Thu, Uijlings, Jasper R. R., and Bernardi, Raffaella. Exploiting language models for visual recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 769–779, 2013b.

Le, Dieu-Thu, Uijlings, Jasper, and Bernardi, Raffaella. *Proceedings of the Third Workshop on Vision and Language*, chapter TUHOI: Trento Universal Human Object Interaction Dataset, pp. 17–24. Dublin City University and the Association for Computational Linguistics, 2014.

Lefever, Els and Hoste, Veronique. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pp. 15–20. Association for Computational Linguistics, 2010.

Lefever, Els and Hoste, Véronique. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pp. 158–166, 2013.

Lesk, Michael. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986*, pp. 24–26, 1986.

Levin, Beth. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, 1993.

Li, Li-Jia and Fei-Fei, Li. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8. IEEE, 2007.

Li, Xirong, Lan, Weiyu, Dong, Jianfeng, and Liu, Hailong. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 271–275. ACM, 2016.

Libovický, Jindrich, Helcl, Jindrich, Tlustý, Marek, Bojar, Ondrej, and Pecina, Pavel. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 646–654, 2016.

Lin, Dekang. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 64–71. Association for Computational Linguistics, 1997.

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, and Zitnick, C. Lawrence. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755, 2014.

Lin, Xiao and Parikh, Devi. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pp. 261–277. Springer, 2016.

Lin, Yuri, Michel, Jean-Baptiste, Aiden, Erez Lieberman, Orwant, Jon, Brockman, Will, and Petrov, Slav. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pp. 169–174. Association for Computational Linguistics, 2012.

Liu, Nian and Han, Junwei. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Trans. Image Processing*, 27(7):3264–3274, 2018.

Loeff, Nicolas, Alm, Cecilia Ovesdotter, and Forsyth, David A. Discriminating image senses by clustering with multimodal features. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, pp. 547–554. Association for Computational Linguistics, 2006.

Lu, Cewu, Krishna, Ranjay, Bernstein, Michael, and Fei-Fei, Li. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pp. 852–869. Springer, 2016a.

Lu, Jiasen, Yang, Jianwei, Batra, Dhruv, and Parikh, Devi. Hierarchical question-image co-attention for visual question answering. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and t, R. Garnet (eds.), *Advances in Neural Information Processing Systems 29*, pp. 289–297. Curran Associates, Inc., 2016b.

Lu, Jiasen, Xiong, Caiming, Parikh, Devi, and Socher, Richard. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3242–3250, 2017.

Luong, Thang, Pham, Hieu, and Manning, Christopher D. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2015.

Ma, Lin, Lu, Zhengdong, Shang, Lifeng, and Li, Hang. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2623–2631, 2015.

Ma, Shugao, Bargal, Sarah Adel, Zhang, Jianming, Sigal, Leonid, and Sclaroff, Stan. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 68:334–345, 2017.

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, Huang, Zhiheng, and Yuille, Alan. Deep captioning with multimodal recurrent neural networks (m-rnn). *International Conference on Learning Representations*, 2015.

Maron, Oded and Lozano-Pérez, Tomás. A framework for multiple-instance learning. In Jordan, M. I., Kearns, M. J., and Solla, S. A. (eds.), *Advances in Neural Information Processing Systems 10*, 1997.

Maron, Oded and Ratan, Aparna Lakshmi. Multiple-instance learning for natural scene classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pp. 341–349, 1998.

Marszalek, Marcin, Laptev, Ivan, and Schmid, Cordelia. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936. IEEE, 2009.

Mathe, Stefan and Sminchisescu, Cristian. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012*, pp. 842–856. Springer, 2012.

Mathe, Stefan and Sminchisescu, Cristian. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*, pp. 1923–1931, 2013.

McCarthy, Diana, Koeling, Rob, Weeds, Julie, and Carroll, John. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 279–286. Association for Computational Linguistics, 2004.

Meyer, Christian M and Gurevych, Iryna. Worth its weight in gold or yet another resource—a comparative study of wiktionary, openthesaurus and germanet. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 38–49. Springer, 2010.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a.

Mikolov, Tomas, Yih, Wen-tau, and Zweig, Geoffrey. Linguistic regularities in continuous space word representations. In *hlt-Naacl*, volume 13, pp. 746–751, 2013b.

Miller, George A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Miller, George A, Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, and Miller, Katherine J. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

Mishra, Ajay, Aloimonos, Yiannis, Fah, Cheong Loong, et al. Active segmentation with fixation. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 468–475. IEEE, 2009.

Miyazaki, Takashi and Shimizu, Nobuyuki. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1780–1790. Association for Computational Linguistics, 2016.

Navigli, Roberto. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

Och, Franz Josef and Ney, Hermann. The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449, 2004.

Pan, Junting, Sayrol, Elisa, Giró i Nieto, Xavier, McGuinness, Kevin, and O'Connor, Noel E. Shallow and deep convolutional networks for saliency prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 598–606, 2016.

Papadopoulos, Dim P, Clarke, Alasdair DF, Keller, Frank, and Ferrari, Vittorio. Training object class detectors from eye tracking data. In *European Conference on Computer Vision*, pp. 361–376. Springer, 2014.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1532–1543, 2014.

Plummer, Bryan A, Wang, Liwei, Cervantes, Chris M, Caicedo, Juan C, Hockenmaier, Julia, and Lazebnik, Svetlana. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Post, Matt and Vilar, David. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1314–1324, 2018.

Pustejovsky, James, Do, Tuan, Kehat, Gitit, and Krishnaswamy, Nikhil. The development of multimodal lexical resources. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pp. 41–47, 2016.

Qiao, Tingting, Dong, Jianfeng, and Xu, Duanqing. Exploring human-like attention supervision in visual question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

Rajendran, Janarthanan, Khapra, Mitesh M., Chandar, Sarath, and Ravindran, Balaraman. Bridge correlational neural networks for multilingual multimodal representation learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 171–181, 2016.

Ramanathan, Vignesh, Li, Congcong, Deng, Jia, Han, Wei, Li, Zhen, Gu, Kunlong, Song, Yang, Bengio, Samy, Rossenberg, Chuck, and Fei-Fei, Li. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1100–1109, 2015.

Rao, B Raja. Partial canonical correlations. *Trabajos de estadistica y de investigación operativa*, 20(2-3):211–219, 1969.

Rashtchian, Cyrus, Young, Peter, Hodosh, Micah, and Hockenmaier, Julia. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147. Association for Computational Linguistics, 2010.

Renninger, Laura Walker, Verghese, Preeti, and Coughlan, James. Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, 7(3):6–6, 2007.

Rodriguez, Mikel D, Ahmed, Javed, and Shah, Mubarak. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

Ronchi, Matteo Ruggero and Perona, Pietro. Describing common human visual actions in images. In *Proceedings of the British Machine Vision Conference (BMVC 2015)*, pp. 52.1–52.12. BMVA Press, September 2015. ISBN 1-901725-53-7.

Rothe, Sascha and Schutze, Hinrich. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural*

Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pp. 1793–1803, 2015.

Rotman, Guy, Vulic, Ivan, and Reichart, Roi. Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael S., Berg, Alexander C., and Li, Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Saenko, Kate and Darrell, Trevor. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 1393–1400, 2008.

Schuler, Karin Kipper. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

Schwenk, Holger. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 228–234, 2018.

Sener, Fadime, Bas, Cagdas, and Ikizler-Cinbis, Nazli. On recognizing actions in still images via multiple features. In *European Conference on Computer Vision*, pp. 263–272. Springer, 2012.

Sennrich, Rico, Haddow, Barry, and Birch, Alexandra. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

Sennrich, Rico, Birch, Alexandra, Currey, Anna, Germann, Ulrich, Haddow, Barry, Heafield, Kenneth, Barone, Antonio Valerio Miceli, and Williams, Philip. The university of edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 389–399, 2017.

Sermanet, Pierre, Eigen, David, Zhang, Xiang, Mathieu, Michaël, Fergus, Rob, and LeCun, Yann. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

Shah, Kashif, Wang, Josiah, and Specia, Lucia. Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 660–665, 2016.

Silberer, Carina and Lapata, Mirella. Learning grounded meaning representations with autoencoders. In *ACL (1)*, pp. 721–732, 2014.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Socher, Richard, Karpathy, Andrej, Le, Quoc V., Manning, Christopher D., and Ng, Andrew Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of Association of Computational Linguistics*, 2:207–218, 2014.

Specia, Lucia, Frank, Stella, Sima'an, Khalil, and Elliott, Desmond. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*, 2016.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Tatler, Benjamin W. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17, 2007.

Van Durme, Benjamin D. *Extracting implicit knowledge from text*. PhD thesis, University of Rochester. Department of Computer Science and Department of Linguistics, 2009.

Vendrov, Ivan, Kiros, Ryan, Fidler, Sanja, and Urtasun, Raquel. Order-embeddings of images and language. *International Conference on Learning Representations*, 2016.

Vig, E, Dorr, M, and Cox, D. Saliency-based space-variant descriptor sampling for action recognition. In *Proceedings of the European Conference on Computer Vision*, 2012.

Vilar, David, Xu, Jia, d'Haro, Luis Fernando, and Ney, Hermann. Error analysis of statistical machine translation output. In *Proceedings of LREC*, pp. 697–702, 2006.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3156–3164, 2015.

Vulic, Ivan, Kiela, Douwe, Clark, Stephen, and Moens, Marie-Francine. Multi-modal representations for improved bilingual lexicon learning. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pp. 188, 2016.

Wang, Weiran, Arora, Raman, Livescu, Karen, and Bilmes, Jeff A. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1083–1092, 2015.

Wieting, John, Mallinson, Jonathan, and Gimpel, Kevin. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 274–285, 2017.

Winkler, Stefan and Subramanian, Ramanathan. Overview of eye tracking datasets. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pp. 212–217. IEEE, 2013.

Wu, Hua and Wang, Haifeng. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181, 2007.

Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V, Norouzi, Mohammad, Macherey, Wolfgang, Krikun, Maxim, Cao, Yuan, Gao, Qin, Macherey, Klaus, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yan, Fei and Mikolajczyk, Krystian. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3441–3450, 2015a.

Yan, Fei and Mikolajczyk, Krystian. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3441–3450, 2015b.

Yang, Zichao, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Smola, Alexander J. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016.

Yao, Bangpeng and Fei-Fei, Li. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 9–16. IEEE, 2010.

Yao, Bangpeng, Jiang, Xiaoye, Khosla, Aditya, Lin, Andy Lai, Guibas, Leonidas, and Fei-Fei, Li. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1331–1338. IEEE, 2011.

Yatskar, Mark, Zettlemoyer, Luke, and Farhadi, Ali. Situation recognition: Visual semantic role labeling for image understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 26-July 1, 2016*, 2016.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78, 2014a.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over

event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78, 2014b.

Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78, 2014c.

Yun, Kiwon, Peng, Yifan, Samaras, Dimitris, Zelinsky, Gregory J, and Berg, Tamara L. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 739–746, 2013.

Zhang, Cha, Platt, John C, and Viola, Paul A. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pp. 1417–1424, 2005.

Zhong, Zhi and Ng, Hwee Tou. It makes sense: A wide-coverage word sense disambiguation system for free text. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pp. 78–83, 2010.

Zhou, Bolei, Lapedriza, Àgata, Xiao, Jianxiong, Torralba, Antonio, and Oliva, Aude. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 487–495, 2014.

Zhou, Bolei, Khosla, Aditya, Lapedriza, Àgata, Oliva, Aude, and Torralba, Antonio. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2921–2929, 2016.